Trustworthy Machine Learning in High Dimensions

by

Simone Bombari

July, 2026

A thesis submitted to the
Graduate School
of the
Institute of Science and Technology Austria
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

Committee in charge:

Matthew Kwan, Chair
Marco Mondelli
Christoph Lampert
Guido Montúfar
Robert D. Nowak

Institute of
Science and
Technology
Austria

The thesis of Simone Bombari, titled *Trustworthy Machine Learning in High Dimensions*, is approved by:

Supervisor: Marco Mondelli, ISTA, Klosterneuburg, Austria

Signature: _____

Committee Member: Christoph Lampert, ISTA, Klosterneuburg, Austria

Signature: _____

Committee Member: Guido Montúfar, UCLA, Los Angeles, USA - MPI, Leipzig, Germany

Signature: _____

Committee Member: Robert D. Nowak, University of Wisconsin–Madison, Madison, USA

Signature: _____

Defense Chair: Matthew Kwan, ISTA, Klosterneuburg, Austria

Signature: _____

Signed page is on file

ISTA Thesis, ISSN: 2663-337X

ISBN: 978-3-99078-091-6

I hereby declare that this thesis is my own work and that it does not contain other people's work without this being so stated; this thesis does not contain my previous work without this being stated, and the bibliography contains all the literature that I used in writing the dissertation.

I accept full responsibility for the content and factual accuracy of this work, including the data and their analysis and presentation, and the text and citation of other work.

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee, and that this thesis has not been submitted for a higher degree to any other university or institution.

I certify that any republication of materials presented in this thesis has been approved by the relevant publishers and co-authors.

Signature: _____

Simone Bombari
July, 2026

Signed page is on file

Abstract

Artificial intelligence and machine learning have undergone an unprecedented evolution in the past decade, motivating a research effort toward a theory able to capture the qualitative behavior of large-scale neural systems. A central puzzle has been the clear benefit of scaling architecture size and overfitting the training set in supervised learning tasks. This evidence, in apparent contradiction with classical statistical learning theory, pushed researchers to develop a new theory capturing the interplay between the algorithmic and architectural bias of training and the specific target function, differently from previous methods rooted in uniform stability. This approach has enabled a grounded understanding of novel learning regimes, typically through formal limits where the number of training samples n, data dimensions d, and model parameters p grow to infinity at different rates.

In this thesis, we follow this approach, focusing on the trustworthiness of high-dimensional models: properties that are difficult to control during training or deployment and often emerge under unpredictable or adversarial conditions. In such settings, it is crucial to formally ensure a priori the reliability of machine learning systems.

First, we study *data memorization*, both as label fitting and as the storage of private information about training samples in trained parameters. We prove that $p = \Omega(n)$ parameters are sufficient for a deep neural network to memorize a generic set of labels, and for a model to memorize spurious features across training data. We then give evidence that $p = \Omega(dn)$ parameters are instead necessary for an adversary to reconstruct the full training set from the trained parameters.

Second, we study *robustness*, both to adversarial perturbations and to distribution shift. We first prove that $p = \Omega(dn)$ parameters can be sufficient for a class of neural networks to overfit the training data while guaranteeing robustness to adversarial perturbations. Then, we focus on spurious correlations learning in high-dimensional regression, studying the effect of the ridge regularization parameter in the proportional regime $n = \Theta(d)$, and connecting it via an equivalence argument to the role of over-parameterization $p = \Omega(n)$ in neural networks. We also investigate the architectural bias of attention-based networks, showing that they are sensitive to the replacement of individual words in an embedded sentence, allowing them to generalize on sentences where the contextual meaning depends on one or few words.

Finally, we study differentially *private optimization* in high-dimensional regimes. We prove that standard private gradient methods do not suffer in the over-parameterized regime $p = \Omega(n)$, challenging the current wisdom based on stability-derived generalization bounds. We then consider linear regression in the proportional regime $n = \Theta(d)$, showing that standard private gradient descent can achieve optimal rates under appropriate hyper-parameter scaling, such as sufficiently small gradient clipping constants, whose role is still debated in practice.

Acknowledgements

I would like to start thanking my PhD advisor Marco Mondelli, for giving me the freedom to find and pursue my research interests, and for his unconditional support and guidance during my PhD journey. I also would like to thank Christoph Lampert, Guido Montúfar and Robert D. Nowak for being part of my thesis committee, and Matthew Kwan for being the chair of my defense.

A special thanks to all the coauthors for the projects included in this thesis: Marco Mondelli, Mohammad Hossein Amani, Shayan Kiyani, Leonardo Iurada, Tatiana Tommasi, Jialei Luo and Inbar Seroussi.

Furthermore, I would like to thank all the researchers that provided an helpful discussion for the purpose of any of the chapter of this thesis: Quynh Nguyen, Mahdi Soltanolkotabi, Adel Javanmard, Grigorios Chrysos, Simone Maria Giancola, Mahyar Jafari Nodeh, Christoph Lampert, Marco Miani, GuanWen Qiu, Peter Súkeník, Yizhe Zhu, Hamed Hassani, Lorenzo Beretta, Clément Rebuffel, Diyuan Wu, Edwige Cyffers, Francesco Pedrotti, Nikita P. Kalinin, Pietro Pelliconi, Roodabeh Safavi, Zhichao Wang, Mahdi Haghifam, Elliot Paquette, Thomas Steinke, Kabir Aladin Verchand and Yichen Wang.

I thank Valentin Roth, Roman Vershynin, and Guangyi Zou for the time they dedicated to proving (independently) that if one considers a vector $v = \mathrm{sign}(z) \in \mathbb{R}^d$, where z is a multivariate Gaussian vector, with covariance having condition number κ, and the sign is taken component-wise, then the sub-Gaussian norm of v is upper bounded by $O(\sqrt{\kappa})$ [Zou and Vershynin, 2026, Roth, 2026]. This result was conjectured by Mohammad Hossein Amani and I during the beginning of this PhD, while working on Chapter 2. In the past years I jokingly mentioned multiple times that I would have been able to graduate only after this result was proven, and I am happy that I can fulfill this silly promise.

I am also grateful to the people that offered help and mentorship during my internship at Amazon Web Services: Alessandro Achille, Stefano Soatto, Aditya Golatkar, Luis Goncalves, Zijian Wang, Yu-Xiang Wang and Giovanni Paolini. Thanks to Clément Rebuffel for his mentorship during my internship at G Research, and thanks to my Google fellowship mentors Thomas Steinke and Ahmad Beirami for the valuable discussions and feedback. Thanks to Mahdi Soltanolkotabi and Yizhe Zhu for hosting me at the University of Southern California, and thanks again to you and to Courtney Paquette for writing recommendation letters for my grant and postdoc applications.

Funding resources. This project was partially supported by the 2019 Lopez-Loreta prize, the European Union (ERC, INF^2, project number 101161364), the Austrian Science Fund (FWF) 10.55776/COE12, and a Google PhD fellowship in machine intelligence. Furthermore, the candidate acknowledges the support from the Scientific Service Units of the Institute of Science and Technology Austria through resources provided by Scientific Computing.

Thanks. I consider myself a very lucky person. Mainly because of all the fantastic people that I have or had the privilege to spend time with, and it is non-negotiable to write a couple of pages dedicated to all those who were by my side during the course of my PhD.

I will start with my advisor, Marco, the guide of this journey. I thank you for your patience, enthusiasm, and positivity that allowed me to go through the toughest initial part of this journey. I thank you for your trust, support, and motivation that made the central part rewarding and enjoyable. I thank you for your efforts and time directed to a successful conclusion of this part of my career. When that smallest eigenvalue was not bounded away from zero, when that proof turned out to be wrong in the middle of the night, when there was still no job offer on the table, I knew that in one way or another everything would have worked out at some point, probably because I knew that you were my strongest ally. I thank you for always being an example of an advisor that cares about doing good and transparent science, that spends hours in writing things in the best possible way, and that works harder for the sake of their students. One motivation behind my choice of pursuing the academic career is to, one day, be that same advisor, and to allow young students to have the same growth opportunity that I had in these years. I really hope this is just the beginning of a much longer collaboration and interaction in the future. As always, the best has yet to come.

I will continue with the rest of the group members. I would like to thank Alex and Kevin, who to me was the golden duo of our group for most of the time I was here, for your inimitable style and for your unbounded kindness. I would like to thank the other PhD students and Postdocs with whom I shared office time, conference travels, gourmet food in cafeteria, workout sessions, and weird conversations about whatever was the topic that day: Al, Yihan, Peter, Diyuan, Amedeo and Filip. I thank Shayan and Christos, for the great time spent in Vienna, Athens, Philly, and London, for the massive dinners, for the protein, for the creatine, for the sweaty workouts. I would like to thank Masani, for being my closest companion during the first two years of my PhD, for always being that person I could have counted on, both in the happy and in the sad moments. I also thank all the amazing people I had the opportunity to share valuable time, long walks, and deep conversations: Ali, Konsti, Ele, Giulia, Annie, and Ece. Finally, I thank the amazing people that joined the group only very recently: Clem, Cristina, Emi, Marek, Pablo. I wish I had more time to spend together, and finishing is not easy also because I know how nice is the group I am leaving behind.

Thanks to my roomies: Marko and Francesco. I shared with you most of the ups and downs of this adventure. The homemade pizzas, the invasive time in each other's room, the wrong objects left in the wrong place at the wrong time (yes Marko, you know what I am talking about...), the quarantine, the pastries from Too Good Too Go, the screams, Shivers, De André, the cleaning conditions, the spinning classes, the mold on the wall, the cozy time on our pretty balcony, the gossips, the secrets, everything. I am so glad that that day on our first year we decided to share an apartment and three years of our life together.

Thanks to the mountains Bubi, Cola, Rosa, Enio. For the biking in Geneva, the secret pizza recipes, the cabin in Trondheim, the shared struggles, the crazy time in Bangkok and our escape from angry taxi drivers. For a memorable summer time in Portugal and in India (congratulations again Rosa and Pearl), the clubbing, the choreographies, our songs, the struggle after the Tibetan street food and at the KFC in Albufeira. I know that I will always be able to rely on you, and your uniqueness. To Mio, Bere, Dario and Fabio, for the ski trips, the conferences, the road trips, for an incredible adventure in Yosemite, for still sticking around after all this time, for the time that awaits ahead in NYC.

Thanks to *IT'S OVER 9000*. We met each other 16 years ago, and for some weird reason we are still together. I see you always less and less, but when we met after months you always make it feel like time froze, and that I never left Crema. For the long calls, for the strange time in Amsterdam, for the fights, for the free gym sessions, and for the craziest trip that we are going to do this summer. Thanks especially to Gabbo and Manuel, for always being on my side, for the nights in Curno, for being the people I can literally tell anything and be myself, for being my best friends.

Thanks to my siblings. To Matte, for carrying the responsibility of being the oldest, for defining my childhood, for the recent joint fitness adventure (which we must get back to). To Robi, for being the calmest, for allowing me to forget everything and just focus on playing when I come back, for being the student that surpassed his master. To Ceci, for being the youngest and only girl, for fighting every day for her spot, for deciding to follow a similar path to mine, which gives me the strength to try to be a good example. Thanks to all of you for being the most precious thing I have, for having taught me that in life you win and lose, but that there is a new game every day.

Grazie ai miei genitori, per avermi dato tutto. Grazie Mamma, per avermi insegnato che non basta fare il minimo, che ci sarà sempre chi è più bravo, che chi non ha testa deve avere gambe, ma che in generale è utile usare la testa. Grazie Papà, per avermi insegnato che anche le cose più serie possono essere prese con leggerezza e divertimento, e che anche le cose più leggere vanno affrontate con serietà e dando il massimo. Grazie a entrambi per avermi insegnato il valore di fare un buon lavoro, che niente è gratis, che bisogna rilassarsi e divertirsi, per non avermi mai imposto o impedito nulla, per le vacanze che ancora organizzate, e per essere l'esempio a cui aspiro ogni giorno.

And finally, thanks to Şeyda, for being first my closest friend when I moved to Vienna, and for becoming my girlfriend along the way. Thank you for being one of my first friends when I arrived at IST, for our long breaks over tea on the bridge of building West, and for being that person I always genuinely loved to spend time with, even after 6 months locked on that hill. Thank you for being by my side during the second year, for the jokes, the memes, and for the time we spent nagging about each other's problems. Thank you for all the great moments, trips, adventures, food, games, and workouts we had since then. Thanks for being an amazing person to live with (together with Octi, Po, Scimmio, Dino, and Mohammad), and thanks for your understanding about my routine and travels. Thanks for being that person I can simply be myself everyday, that I can do big plans with, that is efficient, that is always a peaceful anchor when I am stressed, that does not take anything too seriously, that makes amazing gifts, and that does not need to try to be smart and nice, because she is. Finally, thanks for being here next to me at the end of this journey, and for what still has to come in the next chapter of our lives. I love you.

About the Author

Simone obtained his Bachelor's and Master's degrees in Physics at the University of Pisa in Italy, while being enrolled in the professional Master's program at Scuola Normale Superiore. For his Bachelor's thesis, he was a scientific intern at the Complexity Science Hub in Vienna, and for his Master's thesis, he was a visiting graduate researcher at UCLA, advised by Stefano Soatto. He joined ISTA in September 2020, where he started his affiliation with Professor Marco Mondelli. During his PhD, he was broadly interested in the theoretical foundations of machine learning, with a focus on memorization, robustness, and privacy in high-dimensional settings. He was also an intern at Amazon Web Services and G-Research, and a visiting scholar at the University of Southern California, hosted by Mahdi Soltanolkotabi and Yizhe Zhu. His last two years at ISTA were supported by a Google PhD Fellowship. His work has been published at the machine learning conferences ICLR, ICML, and NeurIPS, and in PNAS.

List of Collaborators and Publications

This thesis is based on the following first-author/equal contribution publications of Simone Bombari. We provide a summary of the contributions of each author, referred to by their initials.

- Simone Bombari, Mohammad Hossein Amani, and Marco Mondelli. Memorization and optimization in deep neural networks with minimum over-parameterization. In *Advances in Neural Information Processing Systems*, 2022b
  - Chapter 2 is fully based on this publication
  - MM initially proposed the problem to SB along with related work, supervised SB, and proposed/made improvements for the formal proofs and suggested additional experiments
  - All authors participated in the discussions which led to the development of the main results
  - SB developed and wrote the majority of technical bulk of the paper
  - SB performed all related numerical simulations
  - All authors contributed to the writing and revising of the final manuscript
- Simone Bombari and Marco Mondelli. How spurious features are memorized: Precise analysis for random and NTK features. In *International Conference on Machine Learning*, 2024a
  - Chapter 3 is fully based on this publication
  - SB initially proposed the problem, MM supervised SB, and proposed/made improvements for the formal proofs and suggested additional experiments
  - All authors participated in the discussions which led to the development of the main results
  - SB developed and wrote the majority of technical bulk of the paper
  - SB performed all related numerical simulations
  - All authors contributed to the writing and revising of the final manuscript
- Leonardo Iurada, Simone Bombari, Tatiana Tommasi, and Marco Mondelli. A law of data reconstruction for random features (and beyond). In *International Conference on Learning Representations*, 2026
  - Chapter 4 is fully based on this publication

- SB initially proposed the problem, LI initially investigated it and proposed the method.
- SB developed and wrote the majority of the theoretical results of the paper. MM supervised SB, and proposed/made improvements for the formal proofs.
- LI developed and performed the majority of the experimental results of the paper. MM and TT supervised LI, and suggested additional experiments.
- All the figures are reproduced with the authors' consent
- All authors participated in the discussions which led to the development of the main results
- All authors contributed to the writing and revising of the final manuscript

- Simone Bombari, Shayan Kiyani, and Marco Mondelli. Beyond the universal law of robustness: Sharper laws for random features and neural tangent kernels. In *International Conference on Machine Learning*, 2023
  - Chapter 5 is fully based on this publication
  - SB initially proposed the problem, MM supervised SB, and proposed/made improvements for the formal proofs and suggested additional experiments
  - All authors participated in the discussions which led to the development of the main results
  - SB developed and wrote the majority of technical bulk of the paper
  - SB performed all related numerical simulations
  - All authors contributed to the writing and revising of the final manuscript

- Simone Bombari and Marco Mondelli. Spurious correlations in high dimensional regression: The roles of regularization, simplicity bias and over-parameterization. In *International Conference on Machine Learning*, 2025b
  - Chapter 6 is fully based on this publication
  - SB initially proposed the problem, MM supervised SB, and proposed/made improvements for the formal proofs and suggested additional experiments
  - All authors participated in the discussions which led to the development of the main results
  - SB developed and wrote the majority of technical bulk of the paper
  - SB performed all related numerical simulations
  - All authors contributed to the writing and revising of the final manuscript

- Simone Bombari and Marco Mondelli. Towards understanding the word sensitivity of attention layers: A study via random features. In *International Conference on Machine Learning*, 2024b
  - Chapter 7 is fully based on this publication
  - SB initially proposed the problem, MM supervised SB, and proposed/made improvements for the formal proofs and suggested additional experiments
  - All authors participated in the discussions which led to the development of the main results

- SB developed and wrote the majority of technical bulk of the paper
- SB performed all related numerical simulations
- All authors contributed to the writing and revising of the final manuscript

- Simone Bombari and Marco Mondelli. Privacy for free in the overparameterized regime. *Proceedings of the National Academy of Sciences*, 122(15):e2423072122, 2025a
  - Chapter 8 is fully based on this publication
  - SB initially proposed the problem, MM supervised SB, and proposed/made improvements for the formal proofs and suggested additional experiments
  - All authors participated in the discussions which led to the development of the main results
  - SB developed and wrote the majority of technical bulk of the paper
  - SB performed all related numerical simulations
  - All authors contributed to the writing and revising of the final manuscript

- Simone Bombari, Jialei Luo, Inbar Seroussi, and Marco Mondelli. High-dimensional private linear regression with optimal rates. *arXiv preprint arXiv:2505.16329*, 2026
  - Chapter 9 is fully based on this publication
  - SB initially proposed the problem to IS
  - SB developed and wrote the majority of technical bulk of the paper, except for Appendix J.3, which was mainly developed by IS
  - JL contributed to the developement of the lower bound
  - MM supervised SB, and proposed/made improvements for the formal proofs and suggested additional experiments
  - All authors participated in the discussions which led to the development of the main results
  - SB performed all related numerical simulations
  - All authors contributed to the writing and revising of the final manuscript

Table of Contents

CHAPTER 1

Introduction

In the last decade, machine learning and deep learning have evolved exceptionally fast, both in terms of performance and of the methodologies used to train and deploy these systems. This evolution was arguably mainly reinforced by two ingredients: *more compute* and *more data*. The first followed from the increasingly widespread availability of graphics processing units and the growing investment in computational resources by leading tech companies and academic institutions, while the latter was achieved through the recent tendency to pre-train foundation models on datasets built essentially from the entire web. Thus, we will start this introduction with the following informal statement.

Modern deep learning owes its success to scale:
very large models trained on very large datasets.

This recipe has led to remarkable performance across a broad range of domains, including image recognition [He et al., 2016, Dosovitskiy et al., 2021], machine translation Bahdanau et al. [2014], Vaswani et al. [2017], image and text generation [Ramesh et al., 2021, OpenAI, 2023, Anil et al., 2023], and progress on scientific frontiers such as biology [Jumper et al., 2021, Abramson et al., 2024], and mathematical reasoning [Trinh et al., 2025, Chervonyi et al., 2025].

This empirical success has in turn motivated an extensive theoretical research effort on questions directly connected to some phenomena observed at scale. Beyond their practical relevance, such phenomena also offer questions of independent interest at the intersection of statistics and optimization. Two informal examples directly inspired by our opening statement are *(i)* how many samples are needed to learn a generic task? And *(ii)* why do larger models perform better?

The first question is central in statistical learning theory, where one focuses on the *sample complexity* of a learning task. The second question is less classical, as the reasons behind the benefits of using larger models are more elusive. In particular, empirical evidence motivated the study of learning models in the *over-parameterized* regime, where the number of trainable parameters p is larger than the number of training data points n. In this regime, the model can fit the training data essentially perfectly, yet it can still generalize well on unseen test data, in apparent contradiction with classical learning theory and the bias-variance tradeoff, where the conventional wisdom suggests training learning models in an *under-parameterized*

regime, i.e. where the number of training data points n is much larger than the number of trainable parameters p, which, for simplicity, can be thought of as the model capacity.

In the following Section 1.1, we briefly review the main tension between the observed generalization properties of over-parameterized models and the predictions of classical learning theory. Then, in Section 1.2, we highlight some properties that a theory of modern deep learning should have in order to agree with the observed empirical evidence, introducing notions such as the *implicit bias* of the training algorithm, *benign overfitting*, and the *linear regime* of neural networks. We then introduce the *random features* (RF) model, as a simple, mathematically tractable model, which will be considered in multiple chapters of this thesis. We will argue for its suitability for analyzing the effects of the number of training samples n, data dimensions d, and number of trainable parameters p growing to infinity at different rates, as well as its limitations for the purpose of studying other problems not directly connected to this thesis.

Our main contribution is not strictly connected to the generalization properties of high-dimensional models, but rather to their *trustworthiness*. Informally, these tend to be properties of learning algorithms that cannot be easily controlled during training and validation, as they generally manifest when the model is deployed out of distribution, or used by an adversarial user trying to manipulate the model itself. In such settings, the model's behavior can be very different from those observed during development, and a theoretically principled, a priori understanding of such behavior is crucial to ensure the reliability of machine learning systems in real-world applications. In particular, we focus on questions connected to *memorization*, *robustness*, and *privacy*, problems that we introduce and motivate in Sections 1.3, 1.4, and 1.5, where we also outline the specific associated results presented in the chapters of this thesis.

1.1 Rethinking generalization

Modern neural networks are often much larger than the amount of training data used to train them [He et al., 2016, Krizhevsky et al., 2017]. Notably, they generalize well at test time even after fitting the training data essentially perfectly, and they tend to perform even better as the number of parameters increases. This phenomenon, where *over-parameterized* neural networks perform better than their *under-parameterized* counterparts, seems to defy the classical bias-variance trade-off, which suggests that increasing the model capacity beyond a certain point should lead to worse generalization [Belkin et al., 2019b].

This tension is highlighted by the influential work of Zhang et al. [2017b], titled *"Understanding Deep Learning Requires Rethinking Generalization"*. There, the following experiment is considered: take a standard image dataset such as CIFAR-10 or ImageNet, and create a copy where the input images are kept fixed, but the training labels are replaced with independent random labels. Standard neural networks trained with gradient methods can fit these random labels almost perfectly, but their resulting test accuracy cannot be better than chance on any distribution since there was no statistical pattern to learn during training. On the other hand, *exactly the same training algorithm* used on the original dataset (with true labels) achieves good performance on the test set, while also achieving zero training error. Thus, the same training algorithm applied to the same input data can either result in a model that generalizes well or one that does not, depending only on the label function. Therefore, generalization bounds that provide guarantees without relying on this function cannot by themselves explain the generalization properties of deep learning models.

For example, let us consider generalization bounds based on uniform stability [Bousquet and Elisseeff, 2002]: an algorithm $\mathcal{A}$ is α-uniformly stable with respect to a loss ℓ if, for any pair of adjacent datasets D and D', we have

$$\sup_{z \in \mathcal{Z}} \mathbb{E}_{\mathcal{A}} \left[\ell(\mathcal{A}(D), z) - \ell(\mathcal{A}(D'), z) \right] \leq \alpha.$$

Here, two adjacent datasets are those that differ in only one sample, the expectation is taken over the randomness of the algorithm, and z is a generic test point, taken from the data space $\mathcal{Z}$. If the loss of misclassifying a datapoint is set (without loss of generality) to 1, then the optimization algorithm considered in Zhang et al. [2017b] is not uniformly stable for any value of α smaller than 1. In fact, the algorithm can fit random labels perfectly, so changing one label in the training set can change the output of the resulting trained model, at least on that same sample. Then, since typical generalization bounds based on uniform stability are additive in α (see, e.g., Theorem 12 in Bousquet and Elisseeff [2002] or Theorem 2.2 in Hardt et al. [2016]), the resulting bound is vacuous.

As another example, consider approaches based on the Rademacher complexity of the class of neural networks [Bartlett and Mendelson, 2002, Koltchinskii and Panchenko, 2002]. More specifically, let $\mathcal{H}$ be the class of predictors that can be described by the previous architecture and training pipeline, and assume for simplicity that the predictors are binary classifiers, with outputs in $\{-1, 1\}$. Denoting with $X = (x_1, \ldots, x_n)$ a fixed set of input data, the empirical Rademacher complexity of $\mathcal{H}$ on X is defined as (see Eq. (1) in Zhang et al. [2017b])

$$\widehat{\mathcal{R}}_X(\mathcal{H}) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(x_i) \right],$$

where $\sigma_1, \ldots, \sigma_n$ are i.i.d. Rademacher random variables. Thus, the Rademacher complexity measures how well functions in $\mathcal{H}$ can correlate with random labels. In the example from Zhang et al. [2017b] discussed above, for every realization of the σ_i-s, the neural network can fit such labels. Then, there exists $f_\sigma \in \mathcal{H}$ such that $f_\sigma(x_i) = \sigma_i$ for all $i \in [n]$, which results in $\widehat{\mathcal{R}}_X(\mathcal{H}) = 1$. Then, since typical generalization bounds based on Rademacher complexity are additive in $\widehat{\mathcal{R}}_X(\mathcal{H})$ (see, e.g., Theorem 5 in Bartlett and Mendelson [2002]), the resulting bound is vacuous.

This state of affairs suggests that standard arguments in classical learning theory structurally mismatch the regime of learning models that generalize well despite overfitting the training set. Hence, a modern theory of deep learning has to rely on more hypotheses than just the complexity of the model function class: these include the data distribution, the target function, and the rule by which the optimization algorithm selects one solution among many. This shifts the paradigm from controlling the worst-case capacity of the architecture toward more specific analyses that are relative to a precise statistical model and to the implicit bias of the training algorithm.

1.2 Towards a theory of deep learning

Inspired by the review of Bartlett et al. [2021], titled *"Deep learning, a statistical viewpoint"*, we identify three ingredients of potential importance for a predictive theory of neural networks:

1. The first one is the *implicit bias* of the training algorithm: in over-parameterized problems that have multiple minimizers of the empirical loss, the specific optimization method

plays a crucial role in selecting a minimizer with certain properties. In linear least squares, for example, gradient descent initialized at zero converges to the minimum Euclidean-norm interpolator; in logistic regression with separable data, gradient descent implicitly maximizes the margin of the solution (see Section 3 in Bartlett et al. [2021]).

2. The second one is *benign overfitting*: in some statistical models, the selected minimizer can be decomposed into a component useful for prediction, and a component that serves to overfit the training samples. These two components separately contribute to the final risk of the algorithm (see, e.g., Lemma 2 in Bartlett et al. [2020]), and benign overfitting happens when the overfitting component has a vanishing effect on it. Notably, this phenomenon is strongly tied to the high-dimensional nature of the data, as it requires lower bounds on the ambient dimension d with respect to the number of samples n (see the discussion in Sections 4.3.1 and 4.3.2 in Bartlett et al. [2021]).

3. The third one is the behavior of the *loss landscape* as networks become over-parameterized: in this regime, gradient methods easily converge to global minimizers of the empirical risk, despite the non-convexity of the loss. This puzzling phenomenon can be partially understood through the lens of the *linearized or lazy regime*, where the network is approximated by its first-order Taylor expansion around initialization and training behaves like a linear model with a fixed feature map [Jacot et al., 2018, Chizat et al., 2019], associated to the *neural tangent kernel* of the network. For a certain class of parameterizations and initializations, this approximation becomes more accurate as networks become wider [Liu et al., 2020], justifying linear models as a simplified proxy to analyze over-parameterized neural networks.

Then, to theoretically study deep learning, it is desirable to consider a mathematical model that can capture the properties described above, while remaining sufficiently simple to admit precise mathematical analysis. A prototypical example is the random features (RF) model [Rahimi and Recht, 2007], defined as

$$f_{\mathrm{RF}}(x, \theta) = \varphi(x)^\top \theta = \phi(Vx)^\top \theta,$$

where $x \in \mathbb{R}^d$ is the input, $V \in \mathbb{R}^{p \times d}$ is a random matrix fixed at initialization, ϕ is a non-linearity applied component-wise, and $\theta \in \mathbb{R}^p$ are the trainable parameters. This model can be viewed as a two-layer neural network where the hidden weights are frozen, and it is therefore a generalized linear model with feature map φ. Then, if one considers the square loss, gradient descent with zero initialization converges to the minimum ℓ_2-norm interpolator, which has a closed form. Besides its simplicity, this model is ideal to study the effects of over-parameterization, since the number of trainable parameters p can be chosen independently of n and d, as opposed to, for example, a standard linear regression model $f_{\mathrm{LR}}(x, \theta) = x^\top \theta$, where the number of parameters is tied to the input dimension $p = d$.

RF models have been extensively studied in the recent theoretical literature [Mei and Montanari, 2022, Mei et al., 2021, Hu et al., 2024], and their test error is well understood for a variety of data distributions and target functions. In particular, *(i)* they have infinitely many empirical risk minimizers in the over-parameterized regime, some of which would not generalize well on unseen data, *(ii)* the minimum norm interpolator has been shown to display benign overfitting, and *(iii)* for some initializations, the solution found by gradient descent when training the neural network $f_{\mathrm{NN}}(x, V, \theta) = \phi(Vx)^\top \theta$, where V is also a trainable parameter, converges to the solution found by the same algorithm applied to an RF model with the same initialization

[Jacot et al., 2018][1]. Furthermore, the RF model has been shown to exhibit other behaviors that are qualitatively similar to those of neural networks, such as the double-descent curve of the test error with respect to p [Belkin et al., 2019b, Nakkiran et al., 2020, Mei and Montanari, 2022]. For this reason, the RF model will be considered in multiple chapters of this thesis, especially when the role of over-parameterization in some context is discussed.

The RF model has, of course, limitations: the feature map $\varphi(x)$ is fixed during training, and thus does not allow for *feature learning*, informally referred to as the phenomenon where the hidden weights significantly adapt to the specific data distribution. Quantitatively, this has consequences for the sample complexity of learning certain target functions, such as polynomials of degree q of one-dimensional projections of the input $x^\top w^*$. In the RF model, assuming uniform spherical data and random features, learning such a target requires on the order of d^q samples [Mei et al., 2021, Hu et al., 2024], while it would need only on the order of d samples if one is allowed to take a single gradient update on the random features V [Ba et al., 2022, Damian et al., 2022, Moniri et al., 2023]. A popular alternative to linear over-parameterized models has been that of neural networks in a *mean-field* regime [Mei et al., 2018, 2019]. In this setting, under gradient flow, the evolution of the parameters is described in terms of the distribution of the neuron weights, but its analysis remains challenging except in highly symmetric settings.

1.3 Memorization

In the context of machine learning, the term *memorization* does not have a precise and unique definition. We identify here two main distinct semantic meanings:

- On the one hand, memorization refers to the ability of a model to fit the (possibly arbitrary) training labels, a notion tightly connected to model capacity. Early results on the question of how many samples can be shattered or fitted by a given class of predictors date back to the classical works of Cover [1965], Baum [1988]. More modern literature approaches the same question from an optimization perspective, i.e. whether a gradient descent method can find an interpolating solution in a sufficiently over-parameterized network. In this context, a rich line of work proved optimization guarantees with progressively smaller lower bounds on the number of parameters [Allen-Zhu et al., 2019, Du et al., 2019b, Soltanolkotabi et al., 2018, Oymak and Soltanolkotabi, 2020, Montanari and Zhong, 2020, Bombari et al., 2022b]. From a different perspective, Feldman [2020], Feldman and Zhang [2020] define memorization as a form of leave-one-out output stability, which is strictly connected to label fitting, and study its positive relation to learning and generalization.
- On the other hand, memorization refers to properties closer to the fields of privacy and security. In this context, it refers to the possibility that information about individual training samples can be recoverable from the trained model. This can be done via membership inference attacks, extraction from generated outputs, or more generally by reconstructing the training inputs from the model parameters [Fredrikson et al., 2015, Carlini et al., 2021, 2023a,b, Cooper and Grimmelmann, 2024, Nasr et al., 2025a, Cooper et al., 2025]. In this sense, memorization goes beyond a model fitting the training labels, as it also requires the model to effectively store the training inputs.

[1] For some initializations, the neural tangent kernel of f_{NN} converges (in some sense) to the kernel induced by the random feature map.

In Chapter 2, we focus on the first notion of memorization, showing that in a fully-connected deep neural network, the number of parameters needed to fit the labels of n random inputs can be as small as $p = \tilde{\Omega}(n)$, where the asymptotic notation $\tilde{\Omega}$ hides poly-logarithmic factors. This is a tight result, as it matches the lower bound of $p \geq n$ that follows from the fact that fitting n labels in general requires at least n degrees of freedom. Our proof technique is inspired by prior work on optimization [Oymak and Soltanolkotabi, 2020], and its main technical contribution involves providing a tight lower bound on the smallest eigenvalue of the empirical neural tangent kernel of the network at initialization, in the regime where the width of the network scales like $\tilde{\Omega}(\sqrt{n})$.

In Chapter 3, we focus on a stylized setting of spurious feature memorization, where each training sample has two components $[x, y]$, with y being uninformative for the task, but nevertheless memorized by the model, and usable at test time to infer information about the corresponding core feature x. Our approach relies on quantifying the leave-one-out stability of over-parameterized generalized linear models, which will turn out to be useful in the later Chapter 8 on private optimization.

In Chapter 4, we focus on the problem of reconstructing the full training dataset from the model parameters. There, we show that for an RF model, data reconstruction becomes possible only if the number of parameters is sufficiently large. In particular, this threshold is significantly different from the one required for label fitting: while the latter is $p = \Omega(n)$ (as discussed in Chapter 2), the former requires $p = \Omega(dn)$, which is much larger if the data are high-dimensional. One intuitive explanation is that reconstructing the full training set implies reconstructing n data points in d dimensions, which requires on the order of dn equations for the problem to not be under-determined. Our proof techniques are similar to the ones used in standard analyses of the RF kernel concentration, but require uniform guarantees in $\mathbb{R}^d$ to suggest that the only solution of our proposed reconstruction algorithm is indeed the original training set. Such uniform guarantees are obtained via packing arguments on the d-dimensional sphere, which translate into an extra factor d on the right hand side of the typical lower bound $p = \Omega(n)$ to achieve RF kernel concentration.

1.4 Robustness

As for memorization, the term *robustness* can have multiple meanings in the context of machine learning. We focus here on the following two:

- A model is said to be *adversarially robust* if its prediction does not significantly change for *any* "small" perturbation of the input. To make this more concrete, one could frame this property in the context of image classification, where each input can be represented as a vector of pixel values. Then, an adversarially robust model should not change its prediction on a test sample for any *adversarial* patch added to the image, as long as the, for example, ℓ_2 norm of the patch is much smaller than the norm of the original image. In practice, it has been observed that standard training methods do not produce adversarially robust models [Szegedy et al., 2014, Goodfellow et al., 2015, Madry et al., 2018].

- A model is said to be *robust to distribution shift* if its predictions do not significantly change when the test distribution differs from the training distribution in some structured way. This can again be made more concrete by considering an image classification

task, where the correct classification function (e.g. boats vs cars) is correlated in the training set with a confounding variable (e.g. the background color), but the same correlation does not hold in the test set. In this case, a model might learn the *spurious correlations* between the confounding variable and the label. This, while potentially improving in-distribution performance, can damage out-of-distribution prediction, with also negative consequences in terms of fairness and reliability [Geirhos et al., 2019, 2020, Xiao et al., 2021, Sagawa et al., 2020a].

The label-fitting viewpoint of memorization offers a convenient entry point into adversarial robustness. If ordinary memorization means that the learning model can interpolate the training labels, adversarial robustness can be informally interpreted as the same model *smoothly* interpolating the same labels. For example, Bubeck et al. [2021] quantifies adversarial robustness as the Lipschitz constant of the model $f(x, \theta)$ with respect to the input x. In Chapter 5, we capture this quantity through the *sensitivity* of the model on a test point x as $|x|_2|\nabla_x f(x, \theta)|_2$, where the initial term is included to make our definition of robustness scale invariant. Then, we show that, with high probability, *(i)* some learning models are non-robust no matter their degree of over-parameterization, *(ii)* others can be robust if their number of parameters is at least $p = \Omega(dn)$, and *(iii)* this second family also includes some appropriately initialized two-layer neural networks, thus resolving in the affirmative a previous conjecture from Bubeck et al. [2020]. Notably, this scaling matches the one discussed in Chapter 4 for data reconstruction. We expect this connection to be not just a matter of coincidence, as both problems are strictly tied to the behavior of input space gradients of the model: their norm characterizes the sensitivity to adversarial perturbations, and they are used to optimize the reconstruction loss that recovers the original training images in Chapter 4.

In Chapter 6, we instead focus on the problem of spurious correlations, where the training data are divided into two components $[x, y]$, where x is the core feature, which is informative for the task, and y is a spurious feature, which is (differently from the setting analyzed in Chapter 3) correlated with x, and therefore with the label. Importantly, the label only depends on the core feature, so that it is conditionally independent of y given x. Then, we initially focus on linear regression, in the high-dimensional regime where the number of training samples is assumed to be proportional to the input dimension $n = \Theta(d)$, and we rely on the asymptotic analysis of Han and Xu [2023] to characterize the test error and the amount of spurious correlations learned by the model as a function of the weight decay parameter. Then, through an equivalence argument between linear regression and RF regression [Goldt et al., 2022], we connect these findings to the impact of over-parameterization for this same problem. Finally, we provide a quantitative interpretation of the *simplicity bias*, according to which learning models tend to learn "easier" patterns, even when they come from confounding variables.

In Chapter 7, we look at a problem reminiscent of input-output robustness in the context of attention-based models. In particular, we ask whether changing an individual token (word) in a sentence can lead to a large change in the attention feature map of a simplified one-layer transformer architecture, and we dub this the *word sensitivity* of the model. We frame this as a desirable property for a language model, as contextual meaning in a sentence is often tied to the presence of a specific word, for instance when a negation determines the correct output of a sentiment classifier. Then, we show that, while the word sensitivity of an RF model vanishes with the length of the context, it remains lower bounded by a constant for randomly initialized attention layers, as long as the embedding dimension is of at least the same order as the context length.

1.5 Privacy

The security viewpoint of memorization allows us to connect to the last topic of this introduction, which is *differential privacy* (DP) [Dwork et al., 2006]. DP is currently the standard formal framework for privacy guarantees in machine learning and data analysis, and it quantifies privacy through numerical parameters, which upper bound the impact that a single data point can have on the output of the algorithm. A common formulation is (ε, δ)-DP, which is satisfied by a randomized algorithm $\mathcal{A}$ if, for every pair of adjacent datasets D, D' differing in one sample, and every subset $S \subseteq \mathbb{R}^p$ of the parameter space, we have

$$\mathbb{P}(\mathcal{A}(D) \in S) \leq e^{\varepsilon}\mathbb{P}(\mathcal{A}(D') \in S) + \delta. \tag{1.1}$$

This guarantee is *worst-case*: it must hold uniformly over *all* adjacent datasets and *all* sets S, and the probability is only taken over the randomness of the algorithm. Smaller values of ε and δ correspond to stronger privacy guarantees, as they imply that the output distribution of the algorithm is less sensitive to changes in the input dataset, becoming fully insensitive in the limit $\varepsilon = \delta = 0$.

This worst-case nature of the definition of DP is in sharp contrast with the typical statistical assumptions used in statistical learning theory, where it is common to assume that the data are sampled i.i.d. from a distribution with convenient concentration properties. DP, by definition, does not rely on such assumptions, and its enforced inequality must hold for any pair of adversarially chosen adjacent datasets. Nevertheless, given a DP algorithm, it is in general meaningful to study its performance on data satisfying distributional assumptions. This allows one to identify privacy-utility trade-offs for given learning tasks, as more stringent privacy requirements (e.g. smaller ε and δ) typically result in lower utility of the algorithm used, and a rich line of work spanning over a decade has investigated this trade-off in various settings [Bassily et al., 2014, 2019, Song et al., 2021b, Varshney et al., 2022, Brown et al., 2024b].

In deep learning, the most common training algorithm that is guaranteed to give a custom level of differential privacy is DP gradient descent [Song et al., 2013, Bassily et al., 2014, Abadi et al., 2016]. This algorithm *clips the per-sample gradients* to control the impact that a single sample can have on the parameters of the model at every iteration, *adds Gaussian noise* to the parameter updates, and limits the number of iterations via *early stopping*, typically degrading the test performance. This procedure adds computational overhead and introduces a layer of complexity with respect to non-private optimization, making tasks such as hyper-parameter tuning more challenging.

Multiple references in the field (see, e.g., Kamath [2020], Yu et al. [2021a], Mehta et al. [2022], Koloskova et al. [2025]) suggest that, since the DP noise is introduced in a p-dimensional parameter space, the over-parameterized regime naturally hinders performance in private optimization, as opposed to non-private optimization, where over-parameterization has previously been discussed as beneficial when overfitting is benign. This is also reflected in the fact that existing upper bounds for private optimization typically become vacuous as the parameter dimension p grows larger than the number of training samples n [Bassily et al., 2019, Song et al., 2021b]. However, these results rely on stability arguments that do not capture the implicit bias of DP gradient descent, nor the specific distribution of the data and the target function, an approach that in Section 1.1 we argued to be doomed in explaining the generalization properties of gradient descent in the over-parameterized regime even in the non-private setting. Then, in Chapter 8, we *rethink generalization in differential privacy*, tightly analyzing the trajectory of DP gradient descent in an over-parameterized RF model,

and showing that the cost of privacy can be negligible in the over-parameterized regime if the hyper-parameters of DP gradient descent are chosen appropriately. In this context, the cost of privacy is measured with respect to the baseline generalization error of the solution found by non-private GD, which is known in our setting to be better than chance [Mei et al., 2021] for a wide set of target functions.

In Chapter 9, we take a complementary route and study the well-established problem of DP linear regression, focusing on the high-dimensional proportional regime $n = \Theta(d)$. Existing nearly optimal DP-GD results for linear regression typically set the clipping constant large enough that clipping almost never occurs [Varshney et al., 2022, Liu et al., 2023b, Brown et al., 2024b], a strategy that we also adopt in Chapter 8. This simplifies the analysis, but it also forces larger private noise (see, e.g. Algorithms 8.1 and 9.1). In contrast, our analysis studies the dynamics when the clipping constant is of the same order as the per-sample gradients and shows that aggressive clipping can be beneficial as long as the learning rate is rescaled accordingly. Our method is inspired by Paquette et al. [2022], Collins-Woodfin et al. [2024a], Marshall et al. [2025], and relies on comparing the risk of DP-GD to the solution of a deterministic system of ordinary differential equations, which in turn allows us to tightly analyze the dynamics of the algorithm even when clipping is aggressive. Then, we show that, in the well-conditioned case, a DP-GD algorithm can achieve the minimax rates for this task if the clipping constant and the learning rate schedule are appropriately chosen. Instead, in the ill-conditioned case, we consider a data covariance spectrum that follows a power-law distribution, and we show that the risk follows a power-law scaling law, highlighting the change in the exponent as a function of the privacy parameter.

CHAPTER 2

Memorization and Optimization in Deep Neural Networks with Minimum Over-parameterization

The Neural Tangent Kernel (NTK) has emerged as a powerful tool to provide memorization, optimization and generalization guarantees in deep neural networks. A line of work has studied the NTK spectrum for two-layer and deep networks with at least a layer with $\Omega(n)$ neurons, n being the number of training samples. Furthermore, there is increasing evidence suggesting that deep networks with sub-linear layer widths are powerful memorizers and optimizers, as long as the number of parameters exceeds the number of samples. Thus, a natural open question is whether the NTK is well conditioned in such a challenging sub-linear setup. In this chapter, we answer this question in the affirmative. Our key technical contribution is a lower bound on the smallest NTK eigenvalue for deep networks with the *minimum possible over-parameterization*: up to logarithmic factors, the number of parameters is $\Omega(n)$ and, hence, the number of neurons is as little as $\Omega(\sqrt{n})$. To showcase the applicability of our NTK bounds, we provide two results concerning memorization capacity and optimization guarantees for gradient descent training.

This chapter is based on the work published at NeurIPS 2022: *Memorization and optimization in deep neural networks with minimum over-parameterization*.

2.1 Introduction

Training a neural network is a non-convex problem that exhibits disconnected local minima [Auer et al., 1996, Safran and Shamir, 2018, Yun et al., 2019a]. Yet, in practice gradient descent (GD) and its variants routinely find solutions with zero training loss [Zhang et al., 2017a]. A framework to understand this phenomenon comes from the Neural Tangent Kernel (NTK). This quantity was introduced in Jacot et al. [2018], where it was proved that, during GD training, the network follows the kernel gradient of the functional cost with respect to the NTK. Furthermore, as the layer widths go large, the NTK converges to a deterministic limit which stays constant during training. Hence, in this infinite-width limit, it suffices that the smallest eigenvalue of the NTK is bounded away from 0 for gradient descent to reach zero loss. Going to finite widths, a recipe to prove GD convergence can be summarized as follows:

show that *(i)* the NTK is well conditioned at initialization, and *(ii)* the NTK has not changed significantly by the time GD has reached zero loss (see e.g. [Oymak and Soltanolkotabi, 2019, Chizat et al., 2019, Bartlett et al., 2021]). A number of papers have exploited this recipe for networks with progressively smaller over-parameterization: two-layer networks [Du et al., 2019b, Oymak and Soltanolkotabi, 2020, Song and Yang, 2020, Wu et al., 2019, Song et al., 2021a], deep networks with polynomially wide layers [Allen-Zhu et al., 2019, Du et al., 2019a, Zou et al., 2020, Zou and Gu, 2019], and deep networks with a single wide layer [Nguyen and Mondelli, 2020, Nguyen, 2021]. Besides optimization, showing that the NTK is well conditioned directly implies a result on memorization capacity [Montanari and Zhong, 2020, Nguyen et al., 2021b], and the smallest eigenvalue of the NTK has also been related to generalization [Arora et al., 2019a].

In Nguyen et al. [2021b], it is shown that, given n training samples, a single layer with $\Omega(n)$ neurons suffices for the NTK to be well conditioned in networks of arbitrary depth. However, there is increasing evidence that, in the challenging setup in which the layer widths are *sub-linear* in n, neural networks still memorize the training data [Bubeck et al., 2020, Yun et al., 2019b, Vershynin, 2020], reach zero loss under GD training in the two-layer setting [Soltanolkotabi et al., 2018, Oymak and Soltanolkotabi, 2020, Montanari and Zhong, 2020], and in the deep case, GD explores a nicely behaved region of the loss landscape [Nguyen et al., 2021a]. This is in agreement with simple back-of-the-envelope calculations on CIFAR-10 and ImageNet: CIFAR-10 has $n = 50000$ images and roughly 10^6 parameters suffice to fit random labels [Zhang et al., 2017a]; furthermore, in order to fit random labels to a subset of $1.2 \cdot 10^6$ ImageNet data points, $2.4 \cdot 10^7$ parameters are enough [Zhang et al., 2017a]. These numbers suggest that having a number of parameters of the same order as the dataset size is much closer to practice than having a number of neurons of that order. We also note that, by counting degrees of freedom or bounding the VC dimension [Bartlett et al., 2019], $\Omega(n)$ parameters (and, therefore, $\Omega(\sqrt{n})$ neurons) are in general necessary to fit n data points. This naturally brings forward the following open question:

Is the NTK well conditioned for deep networks with the minimum possible over-parameterization (i.e., containing $\Omega(n)$ parameters corresponding to $\Omega(\sqrt{n})$ neurons)?

Main contributions. In this chapter, *we settle this open question* for a large class of deep networks. We consider *(i)* a smooth activation function, *(ii)* i.i.d. data satisfying Lipschitz concentration (e.g., data with a Gaussian distribution, uniform on the sphere/hypercube, or obtained via a Generative Adversarial Network), *(iii)* a standard initialization of the weights (e.g., He's or LeCun's initialization), and *(iv)* a loose pyramidal topology in which the layer widths can increase by at most a multiplicative constant as the network gets deep. Then, in Theorem 2.6 we show that the NTK is well conditioned under the *minimum possible over-parameterization* requirement: the number of parameters between the last two layers has to be $\tilde{\Omega}(n)$ or, equivalently, the number of neurons has to be $\tilde{\Omega}(\sqrt{n})$, where $\tilde{\Omega}$ includes extra logarithmic factors. We achieve this goal by giving a lower bound on the smallest eigenvalue of the NTK. This lower bound is tight when all the layer widths are of the same order.

Our NTK bounds open the way towards understanding the behavior of deep networks with minimum over-parameterization. In particular, an immediate consequence of the fact that the NTK is well conditioned is a result on the memorization capacity (Corollary 2.10). Furthermore, by suitably choosing the initialization, we provide convergence guarantees for gradient descent training (Theorem 2.11). Finally, we highlight that, in order to obtain our bounds on the

smallest NTK eigenvalue, we give a number of tight estimates on the ℓ_2 norms of feature vectors and of their centered counterparts, which may be of independent interest.

Proof ideas. To prove Theorem 2.6, we restrict to the kernel $K_{L-2} = J_{L-2}J_{L-2}^\top$, where J_{L-2} is the Jacobian of the output w.r.t. the parameters between the last two hidden layers. This suffices as K_{L-2} is a lower bound on the NTK in the positive semi-definite (PSD) sense. We note that the i-th row ($i \in \{1, \dots, n\}$) of J_{L-2} is given by the Kronecker product $\otimes$ between the feature vector at layer $L-2$ and the backpropagation term from the same layer. One key technical hurdle is to center J_{L-2}, so that its rows have the form $u \otimes v - \mathbb{E}[u \otimes v]$, where $\mathbb{E}[u] = \mathbb{E}[v] = 0$, and all expectations are taken with respect to the (random) training data. To do so, we perform three steps of centering: *(i)* we center the feature vectors (corresponding to u), *(ii)* we center the backpropagation terms (corresponding to v), and *(iii)* we center again the whole row (corresponding to $u \otimes v$). These centering steps are approximate in the sense that the centered matrix is *not necessarily* a lower bound (in the PSD sense) on the original one. However, we are able to control the operator norm of the difference, and show that it scales slower than the smallest eigenvalue of the centered kernel. At this point, we leverage the structure of the rows of the centered Jacobian to bound their sub-exponential norm via a version of the Hanson-Wright inequality for (weakly) correlated random vectors [Adamczak, 2015]. Finally, after providing also a tight estimate on the ℓ_2 norms of such rows, we can exploit a result from Adamczak et al. [2011] to lower bound the smallest singular value of a matrix whose rows are independent random vectors with well controlled sub-exponential and ℓ_2 norms.

Existing work bounds the smallest NTK eigenvalue for networks with two layers [Soltanolkotabi et al., 2018, Montanari and Zhong, 2020], or deep networks with a layer containing $\Omega(n)$ *neurons* [Nguyen et al., 2021b]. In particular, Soltanolkotabi et al. [2018] also exploits the results from Adamczak [2015], Adamczak et al. [2011]. However, the centering of the Jacobian is achieved via a combination of whitening and dropping rows, which appears to be difficult to generalize to deep networks. In contrast, the 3-step centering we described above applies to networks of arbitrary depth L. A different approach is put forward in Montanari and Zhong [2020], and it is based on a decomposition of the kernel via spherical harmonics. This technique allows to obtain the exact limit of the smallest NTK eigenvalue, it has been used to analyze random feature models [Ghorbani et al., 2021, Mei and Montanari, 2022, Ghorbani et al., 2020, Mei et al., 2021] and to obtain generalization bounds for two-layer networks (see again [Montanari and Zhong, 2020]). However, understanding how to carry out such a decomposition in the multi-layer setup is an open problem. Finally, Nguyen et al. [2021b] considers the deep case, and it relates the smallest eigenvalue of the NTK to the smallest singular value of a feature matrix. As feature matrices are full rank only when the number of neurons is $\Omega(n)$, this approach is inherently limited to networks with a linear-width layer.

2.2 Related work

Random matrices in deep learning. The limiting spectra of several random matrices related to neural networks have been the subject of a recent line of work. In particular, the Conjugate Kernel (CK) – namely, the Gram matrix of the features from the last hidden layer – has been studied for models with two layers [Pennington and Worah, 2017, Liao and Couillet, 2018, Louart et al., 2018, Péché, 2019], with a bias term [Adlam et al., 2019, Piccolo and Schröder, 2021], and with multiple layers [Benigni and Péché, 2021]. The Hessian matrix has

been considered in Pennington and Bahri [2017], and the closely related Fisher information matrix in Pennington and Worah [2018]. Using tools from free probability, the input-output Jacobian (as opposed to the parameter-output Jacobian considered in this chapter) of deep networks is studied in Pennington et al. [2018] and the NTK of a two-layer model in Adlam and Pennington [2020]. In Fan and Wang [2020], the spectrum of NTK and CK for deep networks is characterized via an iterated Marchenko-Pastur map. The generalization error has also been studied via the spectrum of suitable random matrices, see Hastie et al. [2022], Mei and Montanari [2022], Liao et al. [2020], Montanari and Zhong [2020] and Section 6 of the review by Bartlett et al. [2021]. Most of the existing results focus on the linear-width asymptotic regime, where the widths of the various layers are linearly proportional. An exception is Wang and Zhu [2021], which focuses on the ultra-wide two-layer case ($N_1 \gg n$).

Smallest eigenvalue of empirical kernels. In the line of work mentioned above, the limiting spectrum of the random matrix is characterized in terms of weak convergence, which does not lead to a direct implication on the behavior of its smallest eigenvalue. In fact, lower bounding the smallest eigenvalue often requires understanding the speed of convergence to the limit spectrum. Quantitative bounds for random Fourier features are obtained in Avron et al. [2017]. In the two-layer setting, the smallest NTK eigenvalue is lower bounded in Soltanolkotabi et al. [2018], Montanari and Zhong [2020], Wang and Zhu [2021], and concentration bounds can also be obtained from Adlam and Pennington [2020], Song and Yang [2020], Hu et al. [2020]. In particular, Montanari and Zhong [2020] establishes a tight bound for two-layer networks with roughly as many parameters as training samples (hence, under minimum over-parameterization), and for a wide class of activations. However, this result still requires that $N_1 = \Omega(d)$. For deep networks, a convergence rate on the NTK can be obtained from Arora et al. [2019b] and potentially from Buchanan et al. [2021], however these results require all the layers to be wide. In Nguyen et al. [2021b], it is proved that a single layer of linear width suffices for the NTK to be well conditioned. Here, Theorem 2.6 provides the first result for deep networks under minimum over-parameterization, therefore allowing for sub-linear layer widths.

Memorization capacity and gradient descent training. For binary classification, the memorization capacity of neural networks was first studied by Cover [Cover, 1965], who solved the single-neuron case. Later, Baum [Baum, 1988] proved that, for two-layer networks, the memorization capacity is lower bounded by the number of parameters, and upper bounds of the same order were given in Kowalczyk [1994], Sakurai [1992]. Recently, tight results for two-layer networks in the regression setting have been provided [Bubeck et al., 2020, Montanari and Zhong, 2020]. In particular, Bubeck et al. [2020] generalizes the construction of Baum [1988], while the memorization result of Montanari and Zhong [2020] comes from a bound on the smallest NTK eigenvalue. This same route is also followed to analyze deep networks in Nguyen et al. [2021b] and in this chapter. Results for multiple layers have been provided in Bartlett et al. [2019], Yun et al. [2019b], Vershynin [2020], Ge et al. [2019]. As concerns efficient algorithms achieving memorization, Daniely [2020a,b] study classification with two-layer networks. A popular line of research has exploited the NTK view to give optimization guarantees on gradient descent. Existing work has focused on the two-layer case [Soltanolkotabi et al., 2018, Du et al., 2019b, Oymak and Soltanolkotabi, 2020, Song and Yang, 2020, Wu et al., 2019, Song et al., 2021a], and optimal results in terms of over-parameterization have been obtained [Montanari and Zhong, 2020]. For deep networks, Du et al. [2019a] assumes a lower bound on the smallest NTK eigenvalue, Allen-Zhu et al. [2019], Zou et al. [2020], Zou and Gu [2019] require all

the layer widths to be (rather large) polynomials in the number of samples, and Nguyen and Mondelli [2020], Nguyen [2021] reduce the over-parameterization to a single layer of linear width. In particular, in these last two papers, the analysis of the NTK is reduced to that of a certain feature matrix (the first one in Nguyen and Mondelli [2020] and the last one in Nguyen [2021]). Hence, these approaches do not seem suitable to tackle networks with sub-linear layer widths[1]. Finally, we remark that milder over-paramerization requirements are sufficient for classification problems. More specifically, it has been proved that polylogarithmic width suffices for two-layer [Ji and Telgarsky, 2019] and deep [Chen et al., 2020b] ReLU networks to converge and generalize. In fact, achieving memorization for regression is harder than for classification, where it suffices to satisfy inequality constraints [Montanari and Zhong, 2020].

2.3 Preliminaries

Neural network setup. We consider an L-layer neural network with feature maps $f_l : \mathbb{R}^d \to \mathbb{R}^{N_l}$ defined for every $x \in \mathbb{R}^d$ as

$$f_l(x) = \begin{cases} x, & l = 0, \\ \phi(W_l^\top f_{l-1}), & l \in [L-1], \\ W_L^\top f_{L-1}, & l = L. \end{cases} \tag{2.1}$$

Here, $W_l \in \mathbb{R}^{N_{l-1} \times N_l}$ is the weight matrix at layer l, ϕ is the activation function and, given an integer n, we use the shorthand $[n] = \{1, \dots, n\}$. We assume that the network has a single output, *i.e.* $N_L = 1$ and $W_L \in \mathbb{R}^{N_{L-1}}$, and for consistency we have $N_0 = d$. Let $g_l : \mathbb{R}^d \to \mathbb{R}^{N_l}$ be the pre-activation feature map so that $f_l(x) = \phi(g_l(x))$ for $l \in [L]$. We define $g_0(x) = f_0(x) = x$. Let $X = [x_1, \dots, x_n]^\top \in \mathbb{R}^{n \times d}$ be the data matrix containing n samples in $\mathbb{R}^d$, $\theta = [\mathrm{vec}(W_1), \dots, \mathrm{vec}(W_L)]$ be the vector of the parameters of the network, and $F_L(\theta) = [f_L(x_1), \dots, f_L(x_n)]^\top$ be the network output. We denote by J the Jacobian of F_L with respect to all the parameters of the network:

$$J = \left[\frac{\partial F_L}{\partial \, \mathrm{vec}(W_1)}, \dots, \frac{\partial F_L}{\partial \, \mathrm{vec}(W_L)} \right] \in \mathbb{R}^{n \times \sum_{l=1}^{L} N_{l-1} N_l}. \tag{2.2}$$

Our key object of interest is the *empirical Neural Tangent Kernel (NTK) Gram matrix*, denoted by $K \in \mathbb{R}^{n \times n}$ and defined as:

$$K = JJ^T = \sum_{l=1}^{L} \left[\frac{\partial F_L}{\partial \, \mathrm{vec}(W_l)} \right] \left[\frac{\partial F_L}{\partial \, \mathrm{vec}(W_l)} \right]^\top. \tag{2.3}$$

In Jacot et al. [2018], it is shown that, as $N_l \to \infty$ for all $l \in [L-1]$, K converges to a deterministic *limit*, which stays constant during gradient descent training. The focus of this chapter is on the *finite-width* behavior of the empirical NTK (2.3). Quantitative bounds for the NTK convergence rate can be obtained from Arora et al. [2019b], Buchanan et al. [2021]. However, these bounds lead to a significant over-parameterization requirement, see the discussion at the end of Section 3 in Nguyen et al. [2021b]. Here, our main result consists in showing that the NTK is well conditioned for a class of neural networks with the *minimum possible over-parameterization*, i.e., $\Omega(n)$ parameter or, equivalently, $\Omega(\sqrt{n})$ neurons.

[1] If the width is sub-linear, then the feature matrix is low-rank and, hence, its smallest eigenvalue is 0.

Weight and data distribution. We consider the following initialization of the weight matrices: $(W_l)_{i,j} \sim_{\text{i.i.d.}} \mathcal{N}(0, \beta_l^2/N_{l-1})$ for $l \in [L-1], i \in [N_{l-1}], j \in [N_l]$, where β_l is a numerical constant independent of the layer widths. This covers the popular cases of He's and LeCun's initialization [Glorot and Bengio, 2010, He et al., 2015, LeCun et al., 2012]. For the last layer, we assume that $(W_L)_i \sim_{\text{i.i.d.}} \mathcal{N}(0, \beta_L^2)$ for $i \in [N_{L-1}]$. Throughout the chapter, we let $(x_1, \ldots, x_n)$ be n i.i.d. samples from the data distribution P_X, which satisfies the conditions below.

Assumption 2.1 (Data scaling)**.** *The data distribution P_X satisfies the following properties:*

1. $\int \|x\|_2 \, dP_X(x) = \Theta(\sqrt{d})$.
2. $\int \|x\|_2^2 \, dP_X(x) = \Theta(d)$.
3. $\int \|x - \int x' \, dP_X(x')\|_2^2 \, dP_X(x) = \Omega(d)$.

Assumption 2.2 (Lipschitz concentration)**.** *The data distribution P_X satisfies the* Lipschitz concentration property*. Namely, there exists an absolute constant $c > 0$ such that, for every Lipschitz continuous function $\varphi : \mathbb{R}^d \to \mathbb{R}$, we have $\mathbb{E}|\varphi(X)| < +\infty$, and for all $t > 0$,*

$$\mathbb{P}\left(\left|\varphi(x) - \int \varphi(x') \, dP_X(x')\right| > t\right) \leq 2e^{-ct^2/\|\varphi\|_{\text{Lip}}^2}.$$

Assumption 2.1 is simply a scaling of the training data points and their centered counterparts. Assumption 2.2 covers a number of important cases, e.g., standard Gaussian distribution [Vershynin, 2018], uniform distributions on the sphere and on the unit (binary or continuous) hypercube [Vershynin, 2018], data produced via a Generative Adversarial Network (GAN)[2] [Seddik et al., 2020], and more generally any distribution satisfying the log-Sobolev inequality with a dimension-independent constant. We also remark that Assumption 2.2 is rather common in the related literature [Nguyen et al., 2021b], or it is even replaced by a stronger requirement (e.g., Gaussian distribution or uniform on the sphere) [Montanari and Zhong, 2020].

Assumption 2.3 (Activation function)**.** *The activation function ϕ satisfies the following properties:*

1. *ϕ is a non-linear (and therefore also non-constant) M-Lipschitz function;*
2. *its derivative ϕ' is a M'-Lipschitz function;*

These requirements are satisfied by common activations, e.g. smoothed ReLU, sigmoid, or $\tanh$.

Assumption 2.4 (Network topology)**.** *The network satisfies a loose pyramidal topology condition, i.e. $N_l = O(N_{l-1})$, for all $l \in [L-1]$.*

A *strict* pyramidal topology (namely, non-increasing layer widths) has been considered in prior work concerning the loss landscape [Nguyen and Hein, 2017, 2018] and gradient descent training [Nguyen and Mondelli, 2020]. Our Assumption 2.4 requires a *loose* pyramidal topology in the sense that, as we go deep, the layer widths are allowed to increase by a constant

[2] By applying a Lipschitz map to a standard Gaussian distribution, the map output satisfies Assumption 2.2.

multiplicative factor. We also note that the widths of neural networks used in practice are often large in the first layers and then start decreasing [Han et al., 2017, Simonyan and Zisserman, 2015].

Assumption 2.5 (Over-parameterization)**.** *We have that*

$$n \cdot \log^8 n = o(N_{L-2}N_{L-1}). \tag{2.4}$$

Furthermore, there exists $\gamma > 0$ *such that*

$$n^\gamma = O\left(N_{L-1}\right). \tag{2.5}$$

Condition (2.4) requires the number of parameters between the last two hidden layers to be linear in the number of samples, up to logarithmic factors. This represents our *key over-parameterization condition*. When all the widths have the same scaling, (2.4) reduces to $n = \tilde{o}(d^2)$. This is satisfied by several "standard" datasets, such as MNIST ($n = 60000$, $d = 784$), CIFAR-10 ($n = 5 \cdot 10^4$, $d = 3 \cdot 32^2$), and ImageNet ($n = 1.4 \cdot 10^7$, $d = 2 \cdot 10^5$). We also note that, if $n \gg d^2$, the NTK is low-rank and $\lambda_{\min}(K) = 0$. Furthermore, Corollary 2.10 and Theorem 2.11 cannot generally hold when $n \gg d^2$, as there are more data points to fit than parameters to help with the fitting. A milder requirement is possible for classification tasks, see the discussion at the end of Section 2.2. The second condition (2.5) is rather mild, as γ can be taken to be arbitrarily small, and it avoids an exponential bottleneck in the last hidden layer.

Notation. The feature matrix at layer l is $F_l = [f_l(x_1), \ldots, f_l(x_n)]^\top \in \mathbb{R}^{n \times N_l}$, and $\Sigma_l(x) = \operatorname{diag}([\phi'(g_{l,j}(x))]_{j=1}^{N_l})$ for $l \in [L-1]$, where $g_{l,j}(x)$ is the pre-activation neuron. Given two matrices $F, B \in \mathbb{R}^{m \times n}$, we denote by $F \circ B$ their Hadamard product, and by $F * B = [(F_{1:} \otimes B_{1:}), \ldots, (F_{m:} \otimes B_{m:})]^T \in \mathbb{R}^{m \times n^2}$ their row-wise Kronecker product (also known as Khatri-Rao product). Given a random vector v, let $\|v\|_{\psi_2}$ and $\|v\|_{\psi_1}$ denote its sub-Gaussian and sub-exponential norm, respectively (see also Appendix B for a detailed definition). Given a matrix A, let $\|A\|_{\text{op}}$ be its operator norm, $\|A\|_F$ its Frobenius norm, $\lambda_{\min}(A)$ its smallest eigenvalue, and $\sigma_{\min}(A)$ its smallest singular value. We denote by $\|\varphi\|_{\text{Lip}}$ the Lipschitz constant of the function φ. All the complexity notations $\Omega(\cdot)$, $O(\cdot)$, $o(\cdot)$ and $\Theta(\cdot)$ are understood for sufficiently large $n, d, N_1, N_2, \ldots, N_{L-1}$. Tildes on such symbols are meant to neglect logarithmic factors.

2.4 NTK bounds with minimum over-parameterization

Our main technical contribution on the smallest eigenvalue of the empirical NTK (2.3) is stated below.

Theorem 2.6 (Smallest NTK eigenvalue under minimum over-parameterization)**.** *Consider an* L*-layer neural network* (2.1)*, where the activation function satisfies Assumption 2.3 and the layer widths satisfy Assumptions 2.4 and 2.5. Let* $\{x_i\}_{i=1}^n \sim_{\text{i.i.d.}} P_X$*, where* P_X *satisfies the Assumptions 2.1-2.2, and let* K *be the empirical NTK Gram matrix* (2.3)*. Assume that the weights of the network are initialized as* $(W_l)_{i,j} \sim_{\text{i.i.d.}} \mathcal{N}(0, \beta_l^2/N_{l-1})$ *for* $l \in [L-1]$*,* $i \in [N_{l-1}]$*,* $j \in [N_l]$ *and* $(W_L)_i \sim_{\text{i.i.d.}} \mathcal{N}(0, \beta_L^2)$ *for* $i \in [N_{L-1}]$*. Then, we have*

$$\lambda_{\min}(K) = \Omega(N_{L-2}N_{L-1}), \tag{2.6}$$

with probability at least $1 - Cne^{-c\log^2 N_{L-1}} - Ce^{-c\log^2 n}$ *over* $(x_i)_{i=1}^n$ *and* $(W_k)_{k=1}^L$*, where* c *and* C *are numerical constants. Moreover, we have that*

$$\lambda_{\min}(K) = O(dN_{L-1}), \tag{2.7}$$

with probability at least $1 - Ce^{-cN_{L-1}}$*, over* $(x_i)_{i=1}^n$ *and* $(W_k)_{k=1}^L$.

This result implies that the NTK is well conditioned for a class of networks in which the number of parameters $N_{L-2}N_{L-1}$ between the last two hidden layers is linear in the number of data samples n, up to logarithmic factors. This means that the total number of neurons of the network can be as little as $\tilde{\Omega}(\sqrt{n})$, which meets the *minimum possible amount of over-parameterization*. The lower bound (2.6) and the upper bound (2.7) match when all the widths (up to layer $L-2$) have the same scaling, i.e., $d = \Theta(N_{L-2})$. Throughout the chapter, we do not explicitly track the dependence of our bounds on L, in the sense that the numerical constants c, C may depend on L.

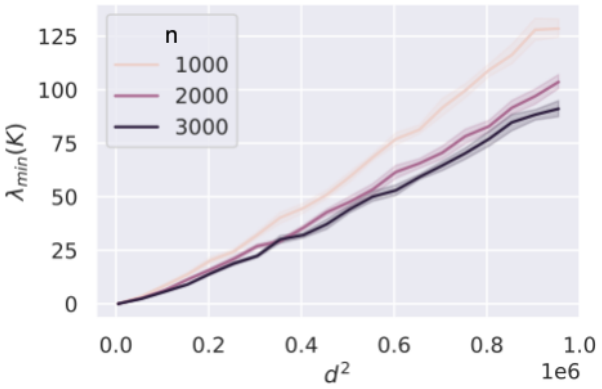


Figure 2.1: $\lambda_{\min}(K)$ in a 3-layer neural network, with sigmoid activation and $d = N_1 = N_2$.

In Figure 2.1, we consider a 3-layer neural network with $d = N_1 = N_2$, and we plot $\lambda_{\min}(K)$ as a function of d^2, for three different values of n. The inputs are sampled from a standard Gaussian distribution, the activation function is the sigmoid $\sigma(x) = (1 + e^{-x})^{-1}$, and we set $\beta_l = 1$ for all $l \in [L]$. We repeat the experiment 10 times, and report average and confidence interval at 1 standard deviation. The linear scaling of $\lambda_{\min}(K)$ in d^2 is in agreement with the result of Theorem 2.6. The code used to obtain the results of Figure 2.1 (and Figure 2.2 as well) is available at `https://github.com/simone-bombari/smallest-eigenvalue-NTK/`.

2.4.1 Roadmap of the proof

After an application of the chain rule and some standard manipulations, we have

$$JJ^\top = \sum_{k=0}^{L-1} K_k, \qquad \text{with } K_k = F_k F_k^\top \circ B_{k+1} B_{k+1}^\top \in \mathbb{R}^{n\times n}, \tag{2.8}$$

where $B_k \in \mathbb{R}^{n\times N_k}$ is a matrix whose i-th row is given by

$$(B_k)_{i:} = \begin{cases} \Sigma_k(x_i)\left(\prod_{l=k+1}^{L-1} W_l \Sigma_l(x_i)\right) W_L, & k \in [L-2], \\ \Sigma_{L-1}(x_i) W_L, & k = L-1, \\ 1, & k = L. \end{cases} \tag{2.9}$$

From the decomposition (2.8), one readily obtains that $K \succeq K_{L-2}$, since K_k is PSD for all $k \in \{0, \dots, L-1\}$. Hence, it suffices to prove a lower bound on $\lambda_{\min}(K_{L-2})$. We denote by J_{L-2} the Jacobian obtained by computing the gradient only over the parameters between

layer $L-2$ and layer $L-1$. Then, $K_{L-2} = J_{L-2}J_{L-2}^\top$ and, by using (2.8) and (2.9), the i-th row of J_{L-2} can be expressed as

$$(J_{L-2})_{i:} = f_{L-2}(x_i) \otimes (D_L \phi'(W_{L-1}^\top f_{L-2}(x_i))), \qquad i \in [n], \tag{2.10}$$

where D_L is a diagonal matrix with the entries of W_L on its diagonal. Note that, if we fix the weight matrices $(W_l)_{l=1}^L$, the rows $((J_{L-2})_{i:})_{i=1}^n$ are i.i.d., as the training data $(x_i)_{i=1}^n$ are i.i.d. too.

The proof consists of two main parts. First, we construct the centered Jacobian $\tilde{J}_{L-2}$, which is obtained from J_{L-2} by iteratively removing expectations with respect to X to its parts, and we show that its smallest singular value is *close* to the smallest singular value of J_{L-2}. The details of this part are contained in Appendix C.3. Second, we bound the smallest eigenvalue of the kernel obtained from the centered Jacobian $\tilde{J}_{L-2}$ by providing an accurate estimate of the ℓ_2 and sub-exponential norms of its rows. The details of this part are contained in Appendix C.4. In order to carry out this program, we exploit a number of concentration results on the ℓ_2 norms of feature and backpropagation vectors, and on the ℓ_2 norms of their centered counterparts. These results are contained in Appendix C.2, and they could be of independent interest. In particular, we provide tight high-probability estimates on *(i)* $\|f_l(x)\|_2$, *(ii)* $\mathbb{E}_x[\|f_l(x)\|_2^2]$, *(iii)* $\mathbb{E}_x[\|f_l(x)\|_2]$, *(iv)* $\mathbb{E}_x[\|f_l(x) - \mathbb{E}_x[f_l(x)]\|_2^2]$, $\mathbb{E}_x[\|f_l(x) - \mathbb{E}_x[f_l(x)]\|_2]$, and $\|f_l(x) - \mathbb{E}_x[f_l(x)]\|_2$, and *(v)* $\|D_L\phi'(g_{L-1}(x)) - \mathbb{E}_x[D_L\phi'(g_{L-1}(x))]\|_2$. Some preliminary calculations are also contained in Appendix C.1.

Part 1: centering. We consider the centered Jacobian $\tilde{J}_{L-2}$, whose i-th row is defined as

$$(\tilde{J}_{L-2})_{i:} = \tilde{f}_{L-2}(x_i) \otimes (D_L\tilde{\phi}'(g_{L-1}(x_i))) - \mathbb{E}_{x_i}\left[\tilde{f}_{L-2}(x_i) \otimes D_L\tilde{\phi}'(g_{L-1}(x_i))\right], \quad i \in [n], \tag{2.11}$$

where $\tilde{f}_{L-2}(x_i) = \tilde{\phi}(g_{L-2}(x_i))$, with $\tilde{\phi}(g_{L-2}(x_i)) = \phi(g_{L-2}(x_i)) - \mathbb{E}_{x_i}[\phi(g_{L-2}(x_i))]$ and $\tilde{\phi}'(g_{L-1}(x_i)) = \phi'(g_{L-1}(x_i)) - \mathbb{E}_{x_i}[\phi'(g_{L-1}(x_i))]$ being the centered versions of the activation function and of its derivative, respectively. Our strategy is to relate $\lambda_{\min}(J_{L-2}J_{L-2}^\top)$ to $\lambda_{\min}(\tilde{J}_{L-2}\tilde{J}_{L-2}^\top)$ via the following two steps.

Step (a): Centering F_{L-2} and B_{L-1}. We show that $\lambda_{\min}(J_{L-2}J_{L-2}^\top) \geq \lambda_{\min}(\tilde{J}_{FB}\tilde{J}_{FB}^\top) - o(N_{L-2}N_{L-1})$, where $\tilde{J}_{FB}\tilde{J}_{FB}^\top = \tilde{F}_{L-2}\tilde{F}_{L-2}^\top \circ \tilde{B}_{L-1}\tilde{B}_{L-1}^\top$ and $\tilde{F}_{L-2} = F_{L-2} - \mathbb{E}_X[F_{L-2}]$ and $\tilde{B}_{L-1} = B_{L-1} - \mathbb{E}_X[B_{L-1}]$ (Lemma C.21 in Appendix C.3.1). Our strategy differs from existing work (e.g., [Soltanolkotabi et al., 2018, Nguyen et al., 2021b]), where either only the backpropagation term is centered (cf. Proposition 7.1 of [Soltanolkotabi et al., 2018]) or only the features are centered (cf. Lemma 5.4 of [Nguyen et al., 2021b]). We remark that, in order to handle deep networks with minimum overparameterization[3], the centering of F_{L-2} and B_{L-1} is crucially performed *at the same time*. In fact, by doing so, certain terms containing an eigenvalue which is negative and large in modulus suitably cancel. As a result, the difference between the original kernel $J_{L-2}J_{L-2}^\top$ and the centered one $\tilde{J}_{FB}\tilde{J}_{FB}^\top$ can be bounded by a matrix whose operator norm is $o(N_{L-2}N_{L-1})$.

Step (b): Centering everything. We show that $\lambda_{\min}(\tilde{J}_{FB}\tilde{J}_{FB}^\top) \geq \lambda_{\min}(\tilde{J}_{L-2}\tilde{J}_{L-2}^\top) - o(N_{L-2}N_{L-1})$, where the rows of $\tilde{J}_{L-2}$ are given by (2.11) (Lemma C.22 in Appendix C.3.2). The idea is

[3] In contrast, [Soltanolkotabi et al., 2018] considers shallow networks, and [Nguyen et al., 2021b] requires the existence of a layer with roughly $\Omega(n)$ neurons.

to decompose the difference $\tilde{J}_{FB}\tilde{J}_{FB}^\top - \tilde{J}_{L-2}\tilde{J}_{L-2}^\top$ into a rank-1 matrix plus a PSD matrix. Then, in order to bound the operator norm of the rank-1 term, we leverage the general version of the Hanson-Wright inequality given by Theorem 2.3 in [Adamczak, 2015] (see the tail bound on quadratic forms that we provide in Lemma C.5). This strategy avoids whitening and dropping rows (as done in [Soltanolkotabi et al., 2018]) and, hence, appears to be better suited to the deep case.

The main result of this part is stated below, and it is proved by combining Lemmas C.21 and C.22.

Theorem 2.7 (Jacobian centering)**.** *Consider the setting of Theorem 2.6, and let the rows of J_{L-2} and $\tilde{J}_{L-2}$ be given by* (2.10) *and* (2.11)*, respectively. Then, we have*

$$\lambda_{\min}\left(J_{L-2}J_{L-2}^\top\right) \geq \lambda_{\min}\left(\tilde{J}_{L-2}\tilde{J}_{L-2}^\top\right) - o(N_{L-2}N_{L-1}), \tag{2.12}$$

with probability at least $1-C\exp(-c\log^2 N_{L-1})-C\exp(-c\log^2 n)$ over $(x_i)_{i=1}^n$ and $(W_k)_{k=1}^L$, where c, C are numerical constants.

Part 2: bounding the centered Jacobian. If we fix the weight matrices $(W_l)_{l=1}^L$, the rows of $\tilde{J}_{L-2}$ are i.i.d. vectors of the form $u\otimes v - \mathbb{E}\left[u\otimes v\right]$, where both u and v are centered. By exploiting this structure, in Appendix C.4.1 we bound the ℓ_2 and sub-exponential norms of these rows. The ℓ_2 norm is bounded in Lemma C.24, and this result relies on the tight estimates on the ℓ_2 norms of centered features and centered backpropagation terms given by Lemma C.19 and C.20, respectively. The sub-exponential norm is bounded in Lemma C.25, and we use again the tail bound of Lemma C.5, which exploits the version of the Hanson-Wright inequality in [Adamczak, 2015]. At this point, the problem is reduced to bounding the smallest eigenvalue of a matrix such that its rows are i.i.d. and we have a good control on their ℓ_2 and sub-exponential norms. This goal can be achieved via the following result, whose proof follows from Theorem 3.2 in [Adamczak et al., 2011] and it is presented for completeness in Appendix C.4.2.

Proposition 2.8. *Let $\tilde{J}$ be a matrix with rows $\tilde{J}_{i:}\in\mathbb{R}^{N_{L-2}N_{L-1}}$, for $i\in[n]$. Assume that $\{\tilde{J}_{i:}\}_{i\in[n]}$ are independent sub-exponential random vectors with $\psi = \max_i \|\tilde{J}_{i:}\|_{\psi_1}$. Let $\eta_{\mathsf{min}} = \min_i \|\tilde{J}_{i:}\|_2$ and $\eta_{\mathsf{max}} = \max_i \|\tilde{J}_{i:}\|_2$. Set $\xi = \psi K + K'$, and $\Delta := C\xi^2 n^{1/4}(N_{L-1}N_{L-2})^{3/4}$. Then, we have that*

$$\lambda_{\min}\left(\tilde{J}\tilde{J}^\top\right) \geq \eta_{\mathsf{min}}^2 - \Delta \tag{2.13}$$

holds with probability at least

$$1-\exp\left(-cK\sqrt{n}\log\left(\frac{2N_{L-1}N_{L-2}}{n}\right)\right) - \mathbb{P}\left(\eta_{\mathit{max}} \geq K'\sqrt{N_{L-1}N_{L-2}}\right), \tag{2.14}$$

where c and C are numerical constants.

The main result of this part is stated below, and it is proved in Appendix C.4.3.

Theorem 2.9 (Bounding the centered Jacobian)**.** *Consider the setting of Theorem 2.6, and let the rows of $\tilde{J}_{L-2}$ be given by* (2.11)*. Then, we have*

$$\lambda_{\min}\left(\tilde{J}_{L-2}\tilde{J}_{L-2}^\top\right) = \Omega(N_{L-2}N_{L-1}), \tag{2.15}$$

with probability at least $1-Ce^{-c\sqrt{n}} - Cne^{-c\log^2 N_{L-1}}$ over $(x_i)_{i=1}^n$ and $(W_k)_{k=1}^L$, where c and C are numerical constants.

Recall that $K \succeq K_{L-2}$, hence $\lambda_{\min}(K) \geq \lambda_{\min}(K_{L-2}) = \lambda_{\min}\left(J_{L-2}J_{L-2}^\top\right)$. Thus, the lower bound (2.6) follows by combining the results from Theorem 2.7 and Theorem 2.9.

Finally, the upper bound (2.7) is rather straightforward. First, we write

$$\lambda_{\min}(K) = \lambda_{\min}\left(JJ^\top\right) \leq (JJ^\top)_{11} = \sum_{l=0}^{L-1} \|(F_l)_{1:}\|_2^2 \, \|(B_{l+1})_{1:}\|_2^2 \,. \tag{2.16}$$

Then, by combining the bound on $\|(F_l)_{1:}\|_2^2$ of Lemma C.16 with a direct calculation for $\|(B_{l+1})_{1:}\|_2^2$, the desired result (2.7) follows. The details are contained in Appendix C.5.

2.5 Two applications: memorization and optimization

Memorization capacity. The fact that the NTK Gram matrix (2.3) is well conditioned readily implies a result on memorization capacity. This was already observed in [Montanari and Zhong, 2020] for two-layer networks and in [Nguyen et al., 2021b] for deep networks with a layer containing $\Omega(n)$ neurons. Here, by using Theorem 2.6, we show that $\Omega(n)$ parameters between the last pair of hidden layers are enough for the network to fit n real-valued labels up to an arbitrarily small error. The result is stated below and, for completeness, we give the proof in Appendix C.6.

Corollary 2.10 (Memorization)**.** *Consider an L-layer neural network* (2.1)*, where the activation function satisfies Assumption 2.3 and the layer widths satisfy Assumptions 2.4 and 2.5. Let $\{x_i\}_{i=1}^n \sim_{\text{i.i.d.}} P_X$, where P_X satisfies the Assumptions 2.1-2.2. Then, it holds that for every $Y \in \mathbb{R}^n$, and for every $\varepsilon > 0$, there exists a set of parameters θ such that*

$$\|F_L(\theta) - Y\|_2 \leq \varepsilon, \tag{2.17}$$

with probability at least $1 - Cne^{-c\log^2 N_{L-1}} - Ce^{-c\log^2 n}$ over $(x_i)_{i=1}^n$, where c and C are numerical constants.

Gradient descent training. Theorem 2.6 has implications on the convergence of gradient descent algorithms. In particular, by choosing carefully the initialization, we show that gradient descent trained on the n samples $\{x_i\}_{i=1}^n$ with labels $Y = (Y_1, \ldots, Y_n)$ converges to zero loss. This is – at the best of our knowledge – the *first result of this kind for deep networks with minimum over-parameterization*.

To define our initialization, let us assume that the $(L-1)$-th layer has an even number of neurons. With a slight abuse of notation, we indicate its width as $2N_{L-1}$. Thus, W_{L-1} can be written in the form $\left[W_{L-1}^{(1)}, W_{L-1}^{(2)}\right]$, where $W_{L-1}^{(1)}, W_{L-1}^{(2)} \in \mathbb{R}^{N_{L-2}\times N_{L-1}}$. Similarly, $W_L \in \mathbb{R}^{2N_{L-1}}$ is the concatenation of $W_L^{(1)}, W_L^{(2)} \in \mathbb{R}^{N_{L-1}}$. Then, we define the initialization $\theta_0 = [\mathrm{vec}(W_{1,0}), \ldots, \mathrm{vec}(W_{L,0})]$ as follows:

$$\begin{aligned}
(W_{l,0})_{i,j} &\sim_{\text{i.i.d.}} \mathcal{N}(0, \beta_l^2/N_{l-1}),\ l \in [L-2],\, i \in [N_{l-1}],\, j \in [N_l], \\
(W_{L-1,0}^{(1)})_{i,j} &\sim_{\text{i.i.d.}} \mathcal{N}(0, \beta_{L-1}^2/N_{L-2}),\ i \in [N_{L-2}],\, j \in [N_{L-1}],\ \text{and } W_{L-1,0}^{(2)} = W_{L-1,0}^{(1)}, \\
(W_{L,0}^{(1)})_i &\sim_{\text{i.i.d.}} \mathcal{N}(0, \beta_L^2\gamma),\ i \in [N_{L-1}],\ \text{and } W_{L,0}^{(2)} = -W_{L,0}^{(1)}.
\end{aligned} \tag{2.18}$$

As usual, $(\beta_l)_{l=1}^L$ are numerical constants independent of the layer widths. In contrast, the quantity γ will be set to a suitably large value depending on the number of training samples and on the layer widths. In words, the initialization is standard for $W_{1,0}, \dots, W_{L-2,0}$ and $W_{L-1,0}^{(1)}$; the variance of $W_{L,0}^{(1)}$ is boosted up by a factor γ; and we duplicate the neurons of layer $L-1$ so that $W_{L-1,0}^{(2)} = W_{L-1,0}^{(1)}$ and $W_{L,0}^{(2)} = -W_{L,0}^{(1)}$. At this point, we are ready to state our optimization result.

Theorem 2.11 (Optimization via gradient descent)**.** *Consider an L-layer neural network* (2.1)*, where the activation function satisfies Assumption 2.3 and the layer widths satisfy Assumptions 2.4 and 2.5. Consider training data $\{x_i\}_{i=1}^n \sim_{\text{i.i.d.}} P_X$, where P_X satisfies the Assumptions 2.1-2.2, with labels $Y \in \mathbb{R}^n$ such that $\|Y\|_2 = \Theta(\sqrt{n})$. We consider solving the least-squares optimization problem*

$$\min_\theta \mathcal{L}(\theta) := \frac{1}{2} \min_\theta \|F_L(\theta) - Y\|_2^2, \tag{2.19}$$

by running gradient descent updates of the form $\theta_{t+1} = \theta_t - \eta \nabla \mathcal{L}(\theta_t)$, where the initialization θ_0 is defined in (2.18) *with $\gamma = d^3 n^2$ and $\eta \le C(\gamma n d N_{L-1})^{-1}$. Then, for all $t \ge 1$,*

$$\mathcal{L}(\theta_t) \le (1 - c\eta\gamma N_{L-2} N_{L-1})^t \, \mathcal{L}(\theta_0), \tag{2.20}$$

with probability at least $1 - C\,n e^{-c \log^2 N_{L-1}} - C e^{-c \log^2 n}$ over $(x_i)_{i=1}^n$ and the initialization θ_0, where c and C are numerical constants.

In words, (2.20) guarantees that gradient descent *linearly* converges to a point with zero loss. We also note that (our upper bound on) the convergence rate can be expressed as a simple function of the learning rate η, the number of training samples n, and the layer widths $\{N_i\}_{i=0}^{L-1}$.

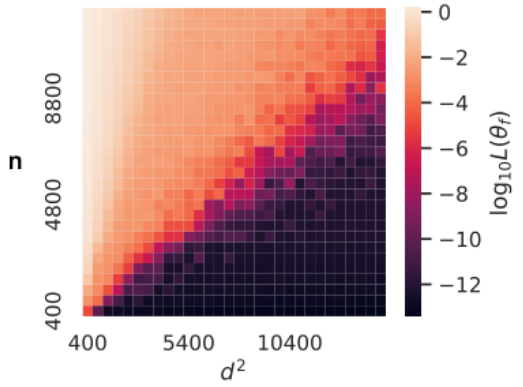


Figure 2.2: Value of the loss $\mathcal{L}(\theta_f)$ once gradient descent has converged for a 4-layer ReLU network with $d = N_1 = N_2 = N_3$.

We now provide a sketch of the argument of Theorem 2.11, deferring the detailed proof to Appendix C.7. As a consequence of Theorem 2.6, the NTK is well conditioned at the initialization θ_0, which implies that gradient descent makes progress at the beginning. The idea is that the NTK remains sufficiently well conditioned also during training and, hence, the loss vanishes with the linear rate in (2.20), since the trained weights remain confined in a ball of sufficiently small radius centered at θ_0. More specifically, we exploit the optimization framework developed in [Oymak and Soltanolkotabi, 2019]. There, a convergence result is proved if the following condition holds: there exists α, β such that

$$\alpha \le \sigma_{\min}(J(\theta)) \le \|J(\theta)\|_{\text{op}} \le \beta, \quad \text{and} \quad \|J(\theta_1) - J(\theta_2)\|_{\text{op}} \le \frac{\alpha^2}{2\beta}, \tag{2.21}$$

for all $\theta, \theta_1, \theta_2 \in \mathcal{D} = \mathcal{B}(\theta_0, R)$, where $\mathcal{D}$ denotes a ball centered at θ_0 of radius $R := 4\|F_L(\theta_0) - Y\|_2/\alpha$ (cf. Proposition C.27). In order to control the radius R, we take advantage of the form (2.18) of the initialization θ_0. In particular, the spectrum of the Jacobian J (and, hence, α) scales linearly in $\sqrt{\gamma}$, and the network output at initialization $F_L(\theta_0)$ is equal to 0

(cf. Lemma C.28). Thus, as $\|Y\|_2 = \Theta(\sqrt{n})$, we have that $R = \Theta(\sqrt{n}/\alpha)$, which can be controlled by choosing suitably γ. At this point, the argument consists in providing a number of estimates on feature vectors, backpropagation terms, and finally the Jacobian $J(\theta)$ for all $\theta \in \mathcal{D}$ (cf. Lemmas C.29-C.35). This allows us to find α, β such that (2.21) holds and conclude.

In Figure 2.2, we give an illustrative example that 4-layer networks achieve 0 loss when the number of parameters is at least linear in the number of training samples, i.e., under minimum over-parameterization. To ease the experimental setup, we use a ReLU activation, with Adam optimizer. We initialize the network as in the setting of Theorem 2.6, picking $\beta_l = 1$ for all $l \in [L]$. The inputs, as well as the targets, are sampled from a standard Gaussian distribution. The plot is averaged over 10 independent trials. As predicted by Theorem 2.11, the loss experiences a phase transition and it goes from 1 to 0 when the layer widths are of order $\sqrt{n}$.

2.6 Concluding remarks

In this chapter, we show that the NTK is well conditioned for deep networks with sub-linear layer widths. As consequences of our NTK bounds, we also prove results on memorization and optimization for such a class of networks. Our approach requires smooth activations and layer widths which do not significantly grow in size with the depth. We believe such assumptions to be purely technical, and we leave as an open question their removal. Our novel NTK analysis could potentially be applied beyond standard gradient descent, e.g., for optimization with momentum [Wang et al., 2021], or federated learning [Huang et al., 2021a].

CHAPTER 3

How Spurious Features Are Memorized: Precise Analysis for Random and NTK Features

Deep learning models are known to overfit and memorize spurious features in the training dataset. While numerous empirical studies have aimed at understanding this phenomenon, a rigorous theoretical framework to quantify it is still missing. In this chapter, we consider spurious features that are uncorrelated with the learning task, and we provide a precise characterization of how they are memorized via two separate terms: *(i)* the *stability* of the model with respect to individual training samples, and *(ii)* the *feature alignment* between the spurious feature and the full sample. While the first term is well established in learning theory and it is connected to the generalization error in classical work, the second one is, to the best of our knowledge, novel. Our key technical result gives a precise characterization of the feature alignment for the two prototypical settings of random features (RF) and neural tangent kernel (NTK) regression. We prove that the memorization of spurious features weakens as the generalization capability increases and, through the analysis of the feature alignment, we unveil the role of the model and of its activation function. Numerical experiments show the predictive power of our theory on standard datasets (MNIST, CIFAR-10).

This chapter is based on the work published at ICML 2024: *How spurious features are memorized: Precise analysis for random and NTK features*.

3.1 Introduction

Neural networks often use features that are not inherently relevant for the intended task. This phenomenon can be caused by positive spurious correlations between certain patterns and the learning task [Geirhos et al., 2020, Xiao et al., 2021], but it occurs even when the patterns are rare [Yang et al., 2022] or simply irrelevant [Hermann and Lampinen, 2020], leading the model to *memorize* spurious relations present in the training data, which are not predictive for the sampling distribution. An extensive empirical effort has aimed at mitigating this phenomenon [Plumb et al., 2022, Chang et al., 2021a]. In fact, the benefits of solving this problem range from robustness to distribution-shift [Geirhos et al., 2019, Zhou et al., 2021a], fairness [Zliobaite, 2015], and data-privacy [Leino and Fredrikson, 2020]. However,

avoiding to overfit spurious features is not always feasible, since memorization can be optimal for accuracy and over-parameterized models often exhibit their best performance when trained long enough to achieve 0 training error [Nakkiran et al., 2020, Feldman, 2020].

In this regard, a related (but separate) body of work has characterized the role of benign overfitting [Belkin, 2021, Bartlett et al., 2020], and it has precisely described the in-distribution generalization of interpolating models, such as random features and neural tangent kernels [Mei and Montanari, 2022, Ghorbani et al., 2021, Montanari and Zhong, 2020]. However, this powerful theoretical machinery does not cover the memorization of spurious features, as noise is generally modelled to be in the labels, rather than in the input data. More generally, while practical work has tried to understand the impact of spurious features and disentangle them from core features in deep learning models [Hermann and Lampinen, 2020, Singla and Feizi, 2022], theoretical approaches remain predominantly directed to understand how learning is impacted by the complexity of the features [Qiu et al., 2023], or the degree of over-parameterization [Sagawa et al., 2020b], without capturing the role of the architecture.

This chapter bridges this gap, offering an analytically tractable framework to understand and quantify the memorization of spurious features. We consider a setting similar to Yang et al. [2022], and we in particular look at the case where the spurious features are not correlated with the true label of the sample (thus the term *memorization*). Formally, we model the sample z as composed by two distinct parts, *i.e.*, $z \equiv [x, y]$, where x is the core feature and y the spurious one, see Figure 3.1 for an illustration. The memorization of spurious features is captured by the correlation between the true label g of the training sample and the output of the model evaluated on the spurious sample $z^s \equiv [-, y]$, where "$-$" corresponds to removing the core feature x (*e.g.*, replacing it with all zeros). In fact, z^s is independent of the label g, as the spurious feature y is un-informative. However, due to memorization, the output of the model evaluated on z^s can still be correlated with g. Our analysis describes quantitatively this phenomenon in the setting of generalized linear regression. Surprisingly, it turns out that the emergence of memorization can be reduced to the separate effect of two distinct components:

1. The *feature alignment* $\mathcal{F}(z^s, z)$, see (3.8). This represents the similarity in feature space between the training sample z and the spurious one z^s; it depends on the feature map of the model and on the rest of the training dataset. To the best of our knowledge, this is the first time that attention is raised over such an object.
2. The *stability* $\mathcal{S}_z$ of the model with respect to z, see Definition 3.1. Similar notions of stability are provided in a rich line of work [Bousquet and Elisseeff, 2002], which relates them to generalization.

Our technical contributions can be summarized as follows:

- We connect the stability in generalized linear regression to the feature alignment between samples, see Lemma 3.2. Then, we show that this connection makes the memorization of spurious features a natural consequence of the generalization error of the model. This is the case when $\mathcal{F}(z^s, z)$ can be well approximated by a constant $\gamma > 0$, independent of the original sample z.
- We focus on two settings widely analyzed in the theoretical literature, *i.e.*, *(i)* random features (RF) [Rahimi and Recht, 2007], and *(ii)* the neural tangent kernel (NTK) [Jacot et al., 2018]. Using tools from high dimensional probability, we prove the concentration

of $\mathcal{F}(z^s, z)$ to a positive constant γ, see Theorems 3.6 and 3.9. For the NTK, we obtain a closed-form expression for γ, which unveils the role of the activation function in the memorization of spurious features.

In a nutshell, our results give a precise characterization of the feature alignment of RF and NTK models. This in turn establishes how the memorization of spurious features grows with the generalization error, and how it depends on the chosen model (RF/NTK) and, in particular, on the activation function. Finally, going beyond RF/NTK models trained on synthetic data, we empirically show that our theoretical predictions transfer to standard datasets (see Figure 3.3 that analyzes the impact of the activation function on MNIST and CIFAR-10) and different neural networks (see Figures 3.4 and 3.5 that consider fully connected, convolutional, and ResNet architectures).

3.2 Related work

Spurious features. Spurious correlations refer to signals that are correlated but not causally related to the learning task [Geirhos et al., 2020, Xiao et al., 2021], and they have been shown to lead to poor out-of-distribution robustness [Geirhos et al., 2019, Zhou et al., 2021a] or biased predictors [Zliobaite, 2015, Seo et al., 2022, Ghosh et al., 2023]. The phenomenon has been studied through the lens of over-parameterization [Sagawa et al., 2020b] and simplicity bias [Hermann and Lampinen, 2020, Shah et al., 2020, Qiu et al., 2023], where the latter refers to models that are inherently prone to learn "easy" patterns first [Kalimeris et al., 2019].

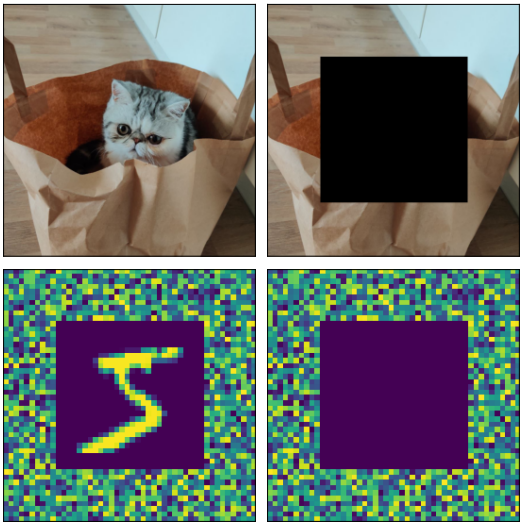


Figure 3.1: Example of a training sample z (top-left) and its spurious counterpart z^s (top-right). In experiments, we add a noise background (y) around the original images (x) before training (bottom-left). We then query the trained model only with the noise component (bottom-right).

This chapter considers spurious features that are independent of the learning task and, hence, focuses on their memorization. Spurious features in the training set can in fact be memorized also if they are irrelevant or rare [Yang et al., 2022, Bansal et al., 2022]. This overfitting can then be used to retrieve information on the training set [Leino and Fredrikson, 2020, Bombari et al., 2022a].

Memorization and stability. Memorization measures the influence of a single sample on the final trained model. Feldman [2020], Feldman and Zhang [2020] point out its advantages on learning from heavy-tailed data, and Arpit et al. [2017], Stephenson et al. [2021] investigate its emergence in neural networks. A related concept is that of leave-one-out stability, which has been studied in a classical line of work: Hoaglin and Welsch [1978] focus on under-parameterized linear models; Bassily et al. [2021], Elisseeff and Pontil [2002], Mukherjee et al. [2006] link it to generalization; and Bousquet and Elisseeff [2002] discuss a wide range of variations on this object.

Random features and neural tangent kernel. The random features (RF) model [Rahimi and Recht, 2007, Pennington and Worah, 2017, Louart et al., 2018] can be regarded as a two-layer neural network with random first layer weights. Its popularity stems from its mathematical tractability and, in contrast with linear regression where the number of parameters equals the input dimension, it captures the statistical effects of over-parameterization, since its number of parameters can scale independently from d and n. Mei and Montanari [2022] characterized the test loss of random features, showing that it displays a double descent curve [Belkin et al., 2019a]. Furthermore, the RF model has been used to understand a wide family of phenomena such as feature learning [Ba et al., 2022, Damian et al., 2022, Moniri et al., 2023], robustness under adversarial attacks [Dohmatob and Bietti, 2022, Bombari et al., 2023, Hassani and Javanmard, 2024], distribution shift [Tripuraneni et al., 2021, Lee et al., 2023, Bombari and Mondelli, 2025b], and scaling laws [Defilippis et al., 2024, Paquette et al., 2024b]. The neural tangent kernel (NTK) can be regarded as the kernel obtained by linearizing a neural network around the initialization [Jacot et al., 2018, Bartlett et al., 2021]. Beyond the discussion of its related literature in Chapter 2, we remark that its behavior has been used in practical work to study adversarial training [Loo et al., 2022] and examples [Tsilivis and Kempe, 2022], and to understand reconstruction attacks for dataset distillation [Loo et al., 2024].

3.3 Preliminaries

Notation. Given a vector v, we denote by $\|v\|_2$ its Euclidean norm. Given $v \in \mathbb{R}^{d_v}$ and $u \in \mathbb{R}^{d_u}$, we denote by $v \otimes u \in \mathbb{R}^{d_v d_u}$ their Kronecker product. Given a matrix $A \in \mathbb{R}^{m \times n}$, we denote by $P_A \in \mathbb{R}^{n \times n}$ the projector over $\mathrm{Span}\{\mathrm{rows}(A)\}$. All the complexity notations $\Omega(\cdot)$, $O(\cdot)$, $o(\cdot)$ and $\Theta(\cdot)$ are understood for sufficiently large data size n, input dimension d, number of neurons k, and number of parameters p. We indicate with $C, c > 0$ numerical constants, independent of n, d, k, p.

Setting. Let (Z, G) be a labelled training dataset, where $Z = [z_1, \dots, z_n]^\top \in \mathbb{R}^{n \times d}$ contains the training data (sampled i.i.d. from a distribution $\mathcal{P}_Z$) on its rows and $G = (g_1, \dots, g_n) \in \mathbb{R}^n$ contains the corresponding labels. We assume the label g_i to be a (eventually noisy) function of the sample z_i. Let $\varphi : \mathbb{R}^d \to \mathbb{R}^p$ be a generic feature map, from the input space to a feature space of dimension p. We consider the following *generalized linear regression* model

$$f(z, \theta) = \varphi(z)^\top \theta, \tag{3.1}$$

where $\varphi(z) \in \mathbb{R}^p$ is the feature vector associated to the input z, and $\theta \in \mathbb{R}^p$ are trainable parameters of the model. We minimize the empirical risk with a quadratic loss:

$$\min_\theta \|\Phi\theta - G\|_2^2, \tag{3.2}$$

where $\Phi := [\varphi(z_1), \dots, \varphi(z_n)]^\top \in \mathbb{R}^{n \times p}$ is the feature matrix. We use the shorthand $K := \Phi\Phi^\top \in \mathbb{R}^{n \times n}$ for the kernel associated with the feature map. If K is invertible (*i.e.*, the model can fit any set of labels G), gradient descent converges to the interpolator which is the closest in ℓ_2 norm to the initialization (see equation (33) in [Bartlett et al., 2021]):

$$\theta^* = \theta_0 + \Phi^+(G - f(Z, \theta_0)), \tag{3.3}$$

where θ^* is the gradient descent solution, θ_0 the initialization, $f(Z, \theta_0) := \Phi\theta_0$ the output of the model (3.1) at initialization, and $\Phi^+ := \Phi^\top K^{-1}$ the Moore-Penrose inverse. Let $z \sim \mathcal{P}_Z$

be an independent test sample. Then, we define the *generalization error* of the trained model as

$$\mathcal{R} = \mathbb{E}_{z\sim\mathcal{P}_Z}\left[(f(z,\theta^*) - g)^2\right], \tag{3.4}$$

where g denotes the ground-truth label of the test sample z.

Stability. Let us introduce quantities related to "incomplete" datasets. We indicate with $\Phi_{-1} \in \mathbb{R}^{(n-1)\times p}$ the feature matrix of the training set *without* the first sample z_1. For simplicity, we focus on the removal of the first sample, and similar considerations hold for the removal of any other sample. In other words, Φ_{-1} is equivalent to Φ, without the first row. Similarly, using (3.3), we indicate with $\theta^*_{-1} := \theta_0 + \Phi^+_{-1}(G_{-1} - f(Z_{-1},\theta_0))$ the set of parameters the algorithm would have converged to if trained over (Z_{-1}, G_{-1}), the original dataset without the first pair sample-label (z_1, g_1). We can now proceed with the definition of our notion of "stability".

Definition 3.1. *Let θ^* (θ^*_{-1}) be the parameters of the model f given by* (3.1) *trained on the dataset Z (Z_{-1}), as in* (3.3)*. We define the* stability $\mathcal{S}_{z_1} : \mathbb{R}^d \to \mathbb{R}$ *with respect to the training sample z_1 as*

$$\mathcal{S}_{z_1} := f(\cdot,\theta^*) - f(\cdot,\theta^*_{-1}). \tag{3.5}$$

This quantity indicates how the trained model changes if we add z_1 to the dataset Z_{-1}. If the training algorithm completely fits the data (as in (3.3)), then $\mathcal{S}_{z_1}(z_1) = g_1 - f(z_1,\theta^*_{-1})$, which implies that

$$\mathbb{E}_{z_1\sim\mathcal{P}_Z}\left[\mathcal{S}^2_{z_1}(z_1)\right] = \mathbb{E}_{z_1\sim\mathcal{P}_Z}\left[\left(f(z_1,\theta^*_{-1}) - g_1\right)^2\right] = \mathbb{E}_{z\sim\mathcal{P}_Z}\left[\left(f(z,\theta^*_{-1}) - g\right)^2\right] =: \mathcal{R}_{Z_{-1}}, \tag{3.6}$$

where the purpose of the second step is just to match the notation used in (3.4), and $\mathcal{R}_{Z_{-1}}$ denotes the generalization error of the algorithm that uses Z_{-1} as training set.

Memorization of spurious features. The input samples are decomposed in two *independent* components, *i.e.*, $z \equiv [x, y]$. With this notation, we mean that $z \in \mathbb{R}^d$ is the concatenation of $x \in \mathbb{R}^{d_x}$ and $y \in \mathbb{R}^{d_y}$ ($d_x + d_y = d$). Here, x is the *core feature* that is useful to accomplish the task (*e.g.*, the cat in top-left image of Figure 3.1), while y is the *spurious feature* containing noise (*e.g.*, the background). Formally, we assume that, for $i \in \{1,\dots,n\}$, $g_i = g(x_i)$, where g is a labelling function, *i.e.*, the label depends only on the core feature x_i and it is independent of the spurious feature y_i (a similar setting was previously considered in Loureiro et al. [2021]). Even if y_i is not useful for learning, the training algorithm may overfit it, memorizing its co-occurrence with the label g_i. This phenomenon could then lead to unpredictable behaviours at test time, such as poor performance on samples $z_t = [x_t, y_i]$ with different information but the same spurious feature [Yang et al., 2022], or data privacy breaches, through extraction of information about x_i via the knowledge of y_i [Leino and Fredrikson, 2020, Bombari et al., 2022a]. Specifically, in computer vision, an unusual background could expose information about the object in the foreground [Leino and Fredrikson, 2020]. In natural language processing, sensitive information (x_i) about an individual (y_i) could be extracted with proper prompting strategies: a language model might in fact be able to successfully auto-complete *"The address of y_i is..."* with *"...x_i"*, as shown by Bombari et al. [2022a] on question-answering tasks. With slight abuse of notation, during the discussion, we will use the term *feature* to indicate both elements in feature space $\varphi(z) \in \mathbb{R}^p$ (as in (3.1)), and portions of the samples x and y. This

choice is common in the related literature, and we will elaborate whenever it could generate confusion.

To quantify the memorization of the spurious feature y_i, we will consider

$$\mathrm{Cov}\left(f(z_i^s, \theta^*), g_i\right), \tag{3.7}$$

which represents the covariance between the true label $g_i = g(x_i)$ and the output of the trained model (3.3) evaluated on $z_i^s = [x, y_i]$, which replaces the core feature x_i with an independent feature x. As x and x_i are independent, this correlation is due to the memorization of y_i. In the experiments, similarly to Singla and Feizi [2022], Yang et al. [2022], we report the *spurious accuracy*, *i.e.*, the fraction of queries $f(z_i^s, \theta^*)$ that returns the correct label $g(x_i)$. For MNIST and CIFAR-10, instead of sampling an independent x, we simply set it to 0, see Figure 3.1. Our code is publicly available at the GitHub repository `simone-bombari/spurious-features-memorization`.

3.4 Memorization and feature alignment

We start by relating the output $f(z_i^s(x), \theta^*)$ evaluated on the spurious sample $z_i^s = [x, y_i]$ with the output $f(z_i, \theta^*)$ evaluated on the original sample z_i. For generalized linear regression, this can be elegantly done via the notion of *stability* of Definition 3.1. By symmetry, from now on, we set $i = 1$ without loss of generality.

Lemma 3.2. *Let $\varphi : \mathbb{R}^d \to \mathbb{R}^p$ be a feature map, such that the induced kernel $K \in \mathbb{R}^{n\times n}$ on the training set is invertible. Let $z_1 \in \mathbb{R}^d$ be an element of the training dataset Z, and $z \in \mathbb{R}^d$ a generic test sample. Let $P_{\Phi_{-1}}$ be the projector over $Span\{\mathrm{rows}(\Phi_{-1})\}$ and $\mathcal{S}_{z_1}$ the stability with respect to z_1, as in Definition 3.1. Define*

$$\mathcal{F}_\varphi(z, z_1) := \frac{\varphi(z)^\top P^\perp_{\Phi_{-1}} \varphi(z_1)}{\left\| P^\perp_{\Phi_{-1}} \varphi(z_1) \right\|_2^2} \tag{3.8}$$

the feature alignment *between z and z_1. Then, we have*

$$\mathcal{S}_{z_1}(z) = \mathcal{F}_\varphi(z, z_1)\, \mathcal{S}_{z_1}(z_1). \tag{3.9}$$

Classical work [Hoaglin and Welsch, 1978] considers a similar problem in the under-parametrized regime, exploiting the projector $H = \Phi(\Phi^\top\Phi)^{-1}\Phi^\top \in \mathbb{R}^{n\times n}$ ($p \le n$ is needed for $\Phi^\top\Phi$ to be invertible), known as the *hat matrix*. In contrast, Lemma 3.2 focuses on the over-parameterized regime and highlights the role of the projector $P_{\Phi_{-1}}$. The different behaviour of under-parameterized and over-parameterized models requires the proof of Lemma 3.2 to follow a different strategy, which is discussed in Appendix D.1.

In words, Lemma 3.2 relates the stability with respect to z_1 evaluated on the two samples z and z_1 through the quantity $\mathcal{F}_\varphi(z, z_1)$, which captures the similarity between z and z_1 in the feature space induced by φ. As a sanity check, the feature alignment between any sample and itself is equal to one, which trivializes (3.9). Then, as z and z_1 become less aligned, the stability $\mathcal{S}_{z_1}(z) = f(z, \theta^*) - f(z, \theta^*_{-1})$ starts to differ from $\mathcal{S}_{z_1}(z_1) = f(z_1, \theta^*) - f(z_1, \theta^*_{-1})$. We note that the feature alignment also depends on the rest of the training set Z_{-1}, as Z_{-1} implicitly appears in the projector $P^\perp_{\Phi_{-1}}$. We also remark that the invertibility of K directly implies that the denominator in (3.8) is different from zero, see Lemma D.1 in Appendix D.1.

Armed with Lemma 3.2, we now characterize the correlation between $f(z_1^s, \theta^*)$ and g_1. Let us replace $\mathcal{F}_\varphi(z_1^s, z_1)$ in (3.9) with a constant $\gamma_\varphi > 0$, independent from z_1. This is justified by Sections 3.5 and 3.6, where we prove the concentration of $\mathcal{F}_\varphi(z_1^s, z_1)$ for the RF and NTK model, respectively. Then, by using the definition of stability in (3.5), we get

$$f(z_1^s, \theta^*) = f(z_1^s, \theta^*_{-1}) + \gamma_\varphi \, \mathcal{S}_{z_1}(z_1) = f(z_1^s, \theta^*_{-1}) + \gamma_\varphi \left(g_1 - f(z_1, \theta^*_{-1})\right). \tag{3.10}$$

Note that $f(z_1^s, \theta^*_{-1})$ is independent from g_1, as it doesn't depend on x_1. In fact, $z_i^s = [x, y_i]$ is independent of x_1, and θ^*_{-1} is not trained on x_1. Thus, if the algorithm is stable, in the sense that $\mathcal{S}_{z_1}(z_1)$ is close to 0, we have that $f(z_1^s, \theta^*)$ has little dependence on g_1. Conversely, if $\mathcal{S}_{z_1}(z_1)$ grows, then $f(z_1^s, \theta^*)$ will start picking up the correlation with g_1. Concretely, we can look at the covariance between $f(z_1^s, \theta^*)$ and g_1, in the probability space of z_1:

$$\begin{aligned} &\mathsf{Cov}\left(f(z_1^s, \theta^*), g_1\right) = \gamma_\varphi \mathsf{Cov}\left(\mathcal{S}_{z_1}(z_1), g_1\right) \\ &\leq \gamma_\varphi \sqrt{\mathsf{Var}\left(\mathcal{S}_{z_1}(z_1)\right) \mathsf{Var}\left(g_1\right)} \leq \gamma_\varphi \sqrt{\mathcal{R}_{Z_{-1}}} \sqrt{\mathsf{Var}\left(g_1\right)}. \end{aligned} \tag{3.11}$$

Here, the first step uses (3.10) and the independence between $f(z_1^s, \theta^*_{-1})$ and g_1, the second step is an application of Cauchy-Schwarz, and the last step follows from (3.6).

The terms γ_φ and $\sqrt{\mathcal{R}_{Z_{-1}}}$ in the RHS of (3.11) show that the memorization of spurious features is affected by two factors: *(i)* the similarity between z_1^s and z_1 (formalized by $\mathcal{F}_\varphi(z_1^s, z_1)$), and *(ii)* the generalization error of the model. While the dependence on the generalization error is not entirely surprising (overfitting causes both the inability of the model to generalize and the memorization of spurious correlations between g_1 and y_1), the key role of the feature alignment (in the form defined in (3.8)) is a-priori far from obvious. Furthermore, via the analysis of $\mathcal{F}_\varphi(z_1^s, z_1)$ in the next two sections, we will show how the memorization is affected by the choice of the feature map and, specifically, of the activation function.

Similarly, we can lower bound the *spurious loss*, defined as

$$\mathcal{L} := \mathbb{E}_{z_1}\left[\left(f(z_1^s, \theta^*) - g_1\right)^2\right]. \tag{3.12}$$

By introducing the shorthands $\bar{f} := \mathbb{E}_{z_1}[f(z_1^s, \theta^*)]$ and $\bar{g} := \mathbb{E}_{z_1}[g_1]$, we have the lower bound:

$$\begin{aligned} \mathcal{L} &= \mathbb{E}_{z_1}\left[\left(\left(f(z_1^s, \theta^*) - \bar{f}\right) - (g_1 - \bar{g}) + \left(\bar{f} - \bar{g}\right)\right)^2\right] \\ &= \mathbb{E}_{z_1}\left[\left(f(z_1^s, \theta^*) - \bar{f}\right)^2\right] + \mathbb{E}_{z_1}\left[(g_1 - \bar{g})^2\right] + \mathbb{E}_{z_1}\left[\left(\bar{f} - \bar{g}\right)^2\right] - 2\,\mathsf{Cov}\left(f(z_1^s, \theta^*), g_1\right) \\ &\geq \mathbb{E}_{z_1}\left[(g_1 - \bar{g})^2\right] - 2\,\gamma_\varphi \sqrt{\mathcal{R}_{Z_{-1}}} \sqrt{\mathsf{Var}\left(g_1\right)}. \end{aligned} \tag{3.13}$$

The first term on the RHS of (3.13) is the minimal loss when no information about x_1 is available (and, thus, the best estimator is the expectation of g_1). The second term indicates how the memorization of the spurious feature y_1 improves the trivial guess $\bar{g}$, and it depends again on $\mathcal{R}_{Z_{-1}}$ and γ_φ.

3.5 Main result for random features

The *random features (RF) model* takes the form

$$f_{\mathsf{RF}}(z, \theta) = \varphi_{\mathsf{RF}}(z)^\top \theta, \qquad \varphi_{\mathsf{RF}}(z) = \phi(Vz), \tag{3.14}$$

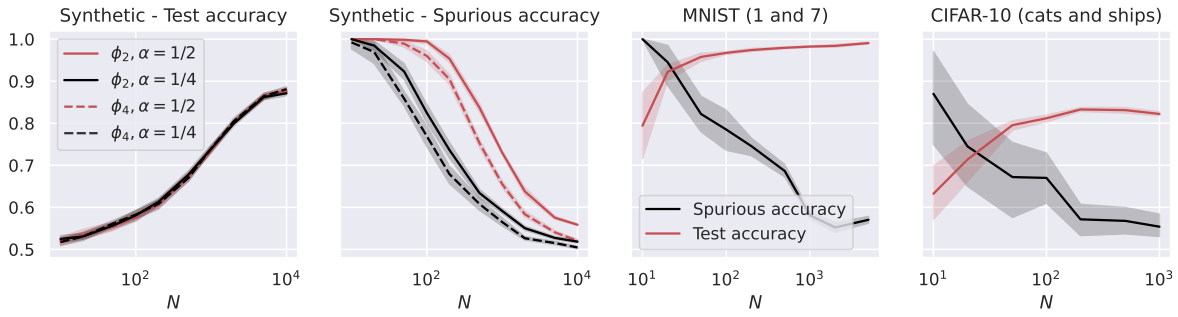


Figure 3.2: Test and spurious accuracies as a function of the number of training samples n, for various binary classification tasks. In the first two plots, we consider the RF model in (3.14) with $k = 10^5$ trained over Gaussian data with $d = 1000$. The labeling function is $g(x) = \text{sign}(u^\top x)$. We repeat the experiments for $\alpha = \{0.25, 0.5\}$ and for the two activations $\phi_2 = h_1 + h_2$ and $\phi_4 = h_1 + h_4$, where h_i denotes the i-th Hermite polynomial (see Appendix B). In the last two plots, we consider the same model with ReLU activation, trained over two MNIST and CIFAR-10 classes. The width of the noise background is 10 pixels for MNIST and 8 pixels for CIFAR-10, see Figure 3.1. The spurious accuracy is obtained by querying the model only with the noise background from the training set, replacing all the other pixels with 0, and taking the sign of the output. As we consider binary classification, an accuracy of 0.5 is achieved by random guessing. We plot the average over 10 independent trials and the confidence band at 1 standard deviation.

where V is a $k \times d$ matrix s.t. $V_{i,j} \sim_{\text{i.i.d.}} \mathcal{N}(0, 1/d)$, and ϕ is an activation applied component-wise. The number of parameters of this model is k, as V is fixed and $\theta \in \mathbb{R}^k$ contains trainable parameters. We denote by μ_l the l-th Hermite coefficient of ϕ (see Appendix B for details).

Assumption 3.3 (Data distribution)**.** *The input data $(z_1, \dots, z_n)$ are n i.i.d. samples from $\mathcal{P}_Z = \mathcal{P}_X \times \mathcal{P}_Y$ s.t. $z_i \in \mathbb{R}^d$ can be written as $z_i = [x_i, y_i]$, with $x_i \in \mathbb{R}^{d_x}$, $y_i \in \mathbb{R}^{d_y}$ and $d = d_x + d_y$. We assume that $x_i \sim \mathcal{P}_X$ is independent of $y_i \sim \mathcal{P}_Y$, and the following holds:*

1. $\|x\|_2 = \sqrt{d_x}$ *and* $\|y\|_2 = \sqrt{d_y}$.
2. $\mathbb{E}[x] = 0$ *and* $\mathbb{E}[y] = 0$.
3. $\mathcal{P}_X$ *and* $\mathcal{P}_Y$ *satisfy* Lipschitz concentration.

The first two assumptions are achieved by data pre-processing and could be relaxed as in Assumption 1 of Bombari et al. [2022b] at the cost of a more involved argument. The third assumption (see Appendix B for details) corresponds to data having well-behaved tails, and it covers a number of important cases, *e.g.*, standard Gaussian [Vershynin, 2018], uniform on the sphere/hypercube [Vershynin, 2018], or data obtained via GANs [Seddik et al., 2020]. This requirement is common in the related literature [Bombari et al., 2022b, Bubeck and Sellke, 2021, Nguyen et al., 2021b] and it is often replaced by a stronger requirement (*e.g.*, data uniform on the sphere), see Montanari and Zhong [2020].

Assumption 3.4 (Over-parameterization and high-dimensional data)**.**

$$n \log^3 n = o(k), \qquad \sqrt{d} \log d = o(k), \qquad k \log^4 k = o(d^2). \tag{3.15}$$

The first condition in (3.15) requires the number of neurons k to scale faster than the number of data points n. This over-parameterization leads to a lower bound on the smallest eigenvalue

of the kernel induced by the feature map, which in turn implies that the model interpolates the data, as required to write (3.3). This over-parameterized regime also achieves minimum test error [Mei and Montanari, 2022]. Combining the second and third conditions in (3.15), we have that k can scale between $\sqrt{d}$ and d^2 (up to $\log$ factors). Finally, merging the first and third condition gives that d^2 scales faster than n. We notice that this holds for standard datasets (MNIST, CIFAR-10 and ImageNet).

Assumption 3.5 (Activation function)**.** *The activation function ϕ is a non-linear L-Lipschitz function.*

Theorem 3.6. *Let Assumptions 3.3, 3.4, and 3.5 hold, and let $x \sim \mathcal{P}_X$ be sampled independently from everything. Consider querying the trained RF model* (3.14) *with $z_1^s = [x, y_1]$. Let $\alpha = d_y/d$ and $\mathcal{F}_{\mathsf{RF}}(z_1^s, z_1)$ be the feature alignment between z_1^s and z_1, as defined in* (3.8)*. Then,*

$$|\mathcal{F}_{\mathsf{RF}}(z_1^s, z_1) - \gamma_{\mathsf{RF}}| = o(1), \tag{3.16}$$

with probability at least $1 - \exp(-c \log^2 n)$ over V, Z and x, where $\gamma_{\mathsf{RF}} \leq 1$ does not depend on z_1 and x. Furthermore, letting μ_l be the l-th Hermite coefficient of the activation ϕ (see Appendix B), we have

$$\gamma_{\mathsf{RF}} > \frac{\sum_{l=2}^{+\infty} \mu_l^2 \alpha^l}{\sum_{l=1}^{+\infty} \mu_l^2} - o(1), \tag{3.17}$$

with probability at least $1 - \exp(-c \log^2 n)$ over V and Z_{-1}, i.e.*, γ_{RF} is bounded away from 0 with high probability.*

Theorem 3.6 proves the concentration of the feature alignment $\mathcal{F}_{\mathsf{RF}}(z_1^s, z_1)$ to a constant γ_{RF} between $\sum_{l=2}^{+\infty} \mu_l^2 \alpha^l / \sum_{l=1}^{+\infty} \mu_l^2 > 0$ and 1. The lower bound increases with α (as expected, since α is the fraction of the input given by the spurious feature), and it depends in a non-trivial way on the activation via its Hermite coefficients. This result validates the argument in (3.10), where we replaced the feature alignment $\mathcal{F}_{\mathsf{RF}}(z_1^s, z_1)$ with a constant $\gamma_{\mathsf{RF}} > 0$. Thus, (3.11) now reads

$$\mathsf{Cov}\left(f_{\mathsf{RF}}(z_1^s, \theta^*), g_1\right) \leq \gamma_{\mathsf{RF}} \sqrt{\mathcal{R}_{Z_{-1}}} \sqrt{\mathsf{Var}\left(g_1\right)}, \tag{3.18}$$

which quantifies the memorization of spurious features in terms of the generalization error and the constant γ_{RF}.

These effects are clearly displayed in Figure 3.2 for binary classification on synthetic (first two plots) and image (last two plots) datasets. Specifically, as the number of samples n increases, the test accuracy increases and the spurious accuracy (obtained by querying the trained model with the spurious sample $f(z_i^s, \theta^*)$) decreases. Furthermore, for the synthetic dataset, while the test accuracy does not depend on α or the activation function, the spurious accuracy increases with α and by taking an activation function with dominant low-order Hermite coefficients, as predicted by (3.17). For an additional experiment highlighting the dependence on α, we refer the reader to Figure D.2 in Appendix D.5.

Proof sketch. We set

$$\gamma_{\mathsf{RF}} := \frac{\mathbb{E}_{z_1, z_1^s}[\varphi_{\mathsf{RF}}(z_1^s)^\top P^{\perp}_{\Phi_{-1}} \varphi_{\mathsf{RF}}(z_1)]}{\mathbb{E}_{z_1}[\|P^{\perp}_{\Phi_{-1}} \varphi_{\mathsf{RF}}(z_1)\|_2^2]}. \tag{3.19}$$

Here, $P_{\Phi_{-1}}$ is the projector over $\mathrm{Span}\{\mathrm{rows}(\Phi_{\mathsf{RF},-1})\}$ and $\Phi_{\mathsf{RF},-1}$ the RF feature matrix after removing the first row. With this choice, the numerator and denominator of γ_{RF} equal the

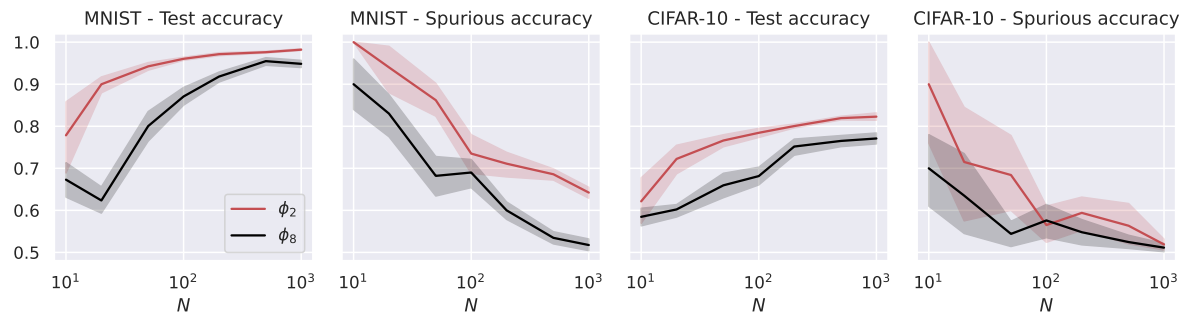


Figure 3.3: We consider the NTK model in (3.24) with $k = 100$, trained on MNIST (digits 1 and 7, first and second plots), and CIFAR-10 (cats and ships, third and fourth plots). We repeat the experiments for activations whose derivatives are $\phi_2' = h_0 + h_1$ and $\phi_8' = h_0 + h_7$, where h_i denotes the i-th Hermite polynomial (see Appendix B). The rest of the setup is the same as that of Figure 3.2.

expectations of the corresponding quantities appearing in $\mathcal{F}_{\mathsf{RF}}(z_1^s, z_1)$. Thus, the concentration result in (3.16) is obtained from the general form of the Hanson-Wright inequality in [Adamczak, 2015], see Lemma D.12. The upper bound $\gamma_{\mathsf{RF}} \leq 1$ follows from an application of Cauchy-Schwarz inequality. In contrast, the lower bound is more involved and it is obtained via the three steps below.

Step 1: Centering the feature map φ_{RF}. We extract the term $\mathbb{E}_V[\phi(Vz)]$ from the expression of $\mathcal{F}_{\mathsf{RF}}$ and show it can be neglected, due to the specific structure of $P_{\Phi_{-1}}^{\perp}$. Specifically, letting $\tilde{\varphi}_{\mathsf{RF}}(z) := \varphi_{\mathsf{RF}}(z) - \mathbb{E}_V[\varphi_{\mathsf{RF}}(z)]$, we have

$$\mathcal{F}_{\mathsf{RF}}(z_1^s, z_1) \simeq \frac{\tilde{\varphi}_{\mathsf{RF}}(z_1^s)^\top P_{\Phi_{-1}}^{\perp} \tilde{\varphi}_{\mathsf{RF}}(z_1)}{\left\| P_{\Phi_{-1}}^{\perp} \tilde{\varphi}_{\mathsf{RF}}(z_1) \right\|_2^2} = \frac{\tilde{\varphi}_{\mathsf{RF}}(z_1^s)^\top \tilde{\varphi}_{\mathsf{RF}}(z_1) - \tilde{\varphi}_{\mathsf{RF}}(z_1^s)^\top P_{\Phi_{-1}} \tilde{\varphi}_{\mathsf{RF}}(z_1)}{\left\| P_{\Phi_{-1}}^{\perp} \tilde{\varphi}_{\mathsf{RF}}(z_1) \right\|_2^2}, \quad (3.20)$$

where $\simeq$ denotes an equality up to a $o(1)$ term. This is formalized in Lemma D.8.

Step 2: Linearization of the centered feature map $\tilde{\varphi}_{\mathsf{RF}}$. We consider the terms $\tilde{\varphi}_{\mathsf{RF}}(z_1^s), \tilde{\varphi}_{\mathsf{RF}}(z_1)$ that multiply $P_{\Phi_{-1}}$ in the RHS of (3.20), and we show that they are well approximated by their first-order Hermite expansions ($\mu_1 V z_1^s$ and $\mu_1 V z_1$, respectively). In fact, the rest of the Hermite series scales at most as n/d^2, which is negligible due to Assumption 3.4. Specifically, Lemma D.9 implies

$$\frac{\tilde{\varphi}_{\mathsf{RF}}(z_1^s)^\top \tilde{\varphi}_{\mathsf{RF}}(z_1) - \tilde{\varphi}_{\mathsf{RF}}(z_1^s)^\top P_{\Phi_{-1}} \tilde{\varphi}_{\mathsf{RF}}(z_1)}{\left\| P_{\Phi_{-1}}^{\perp} \tilde{\varphi}_{\mathsf{RF}}(z_1) \right\|_2^2} \simeq \frac{\tilde{\varphi}_{\mathsf{RF}}(z_1^s)^\top \tilde{\varphi}_{\mathsf{RF}}(z_1) - \mu_1^2 (V z_1^s)^\top P_{\Phi_{-1}} (V z_1)}{\left\| P_{\Phi_{-1}}^{\perp} \tilde{\varphi}_{\mathsf{RF}}(z_1) \right\|_2^2}. \quad (3.21)$$

Step 3: Lower bound in terms of α and $\{\mu_l\}_{l \geq 2}$. To conclude, we express the RHS of (3.21) as follows:

$$\begin{aligned} &\frac{\tilde{\varphi}_{\mathsf{RF}}(z_1^s)^\top \tilde{\varphi}_{\mathsf{RF}}(z_1) - \mu_1^2 (V z_1^s)^\top (V z_1) + \mu_1^2 (V z_1^s)^\top P_{\Phi_{-1}}^{\perp} (V z_1)}{\|\tilde{\varphi}_{\mathsf{RF}}(z_1)\|_2^2 - \left\| P_{\Phi_{-1}} \tilde{\varphi}_{\mathsf{RF}}(z_1) \right\|_2^2} \\ &\gtrsim \frac{\tilde{\varphi}_{\mathsf{RF}}(z_1^s)^\top \tilde{\varphi}_{\mathsf{RF}}(z_1) - \mu_1^2 (V z_1^s)^\top (V z_1)}{\|\tilde{\varphi}_{\mathsf{RF}}(z_1)\|_2^2} \simeq \frac{\sum_{l=2}^{+\infty} \mu_l^2 \alpha^l}{\sum_{l=1}^{+\infty} \mu_l^2} > 0, \end{aligned} \quad (3.22)$$

where $\gtrsim$ denotes an inequality up to a $o(1)$ term. As $P_{\Phi_{-1}} = I - P_{\Phi_{-1}}^{\perp}$, the term in the first line equals the RHS of (3.21). Next, we show that $\mu_1^2 (V z_1^s)^\top P_{\Phi_{-1}}^{\perp} (V z_1)$ is equal to

$\mu_1^2(V[y_1,0])^\top P_{\Phi_{-1}}^\perp(V[y_1,0])$ (which corresponds to the common noise part in z_1, z_1^s) plus a vanishing term, see Lemma D.10. As $\mu_1^2(V[y_1,0])^\top P_{\Phi_{-1}}^\perp(V[y_1,0]) \geq 0$, the inequality in the second line follows. The last step is obtained by showing concentration over V of numerator and denominator. The expression on the RHS of (3.22) is strictly positive as $\alpha > 0$ and ϕ is non-linear by Assumption 3.5.

3.6 Main result for NTK features

We consider the following two-layer neural network

$$f_{\mathsf{nn}}(z,w) = \sum_{i=1}^{k} \phi\left(W_{i:}z\right). \tag{3.23}$$

The hidden layer contains k neurons; ϕ is an activation function applied component-wise; $W \in \mathbb{R}^{k\times d}$ denotes the weights of the hidden layer; $W_{i:}$ denotes the i-th row of W; and we set the k weights of the second layer to 1. We indicate with w the vector containing the parameters of this model, *i.e.*, $w = [\mathsf{vec}(W)] \in \mathbb{R}^p$, with $p = kd$. We initialize the network with standard (*e.g.*, He's or LeCun's) initialization, *i.e.*, $[W_0]_{i,j} \sim_{\text{i.i.d.}} \mathcal{N}(0, 1/d)$.

The *NTK regression model* takes the form

$$f_{\mathsf{NTK}}(z,\theta) = \varphi_{\mathsf{NTK}}(z)^\top\theta, \qquad \varphi_{\mathsf{NTK}}(z) = \nabla_w f_{\mathsf{nn}}(z,w)|_{w=w_0} = z \otimes \phi'(W_0 z), \tag{3.24}$$

where $\nabla_w f_{\mathsf{nn}}$ in the second line is computed via the chain rule. The vector of trainable parameters is $\theta \in \mathbb{R}^p$, with $p = kd$, which is initialized with $\theta_0 = w_0 = [\mathsf{vec}(W_0)]$. We note that $f_{\mathsf{NTK}}(z,\theta)$ is the linearization of $f_{\mathsf{nn}}(z,w)$ around the initial point w_0 [Bartlett et al., 2021, Jacot et al., 2018], and the model in (3.24) corresponds to training the two-layer network (3.23) in the lazy regime [Chizat et al., 2019, Oymak and Soltanolkotabi, 2019]. As such, it has received significant attention in the theoretical literature, see *e.g.* [Bombari et al., 2023, Dohmatob and Bietti, 2022, Montanari and Zhong, 2020].

Assumption 3.7 (Over-parameterization and topology)**.**

$$n\log^8 n = o(kd), \qquad n > d, \qquad k = O\left(d\right). \tag{3.25}$$

The first condition is the smallest (up to $\log$ factors) over-parameterization that guarantees interpolation [Bombari et al., 2022b]. The second condition is rather mild (it is easily satisfied by standard datasets) and purely technical. The third condition is required to lower bound the smallest eigenvalue of the kernel induced by the feature map (3.24), and a stronger requirement, *i.e.*, the strict inequality $k < d$, has appeared in prior work [Nguyen and Hein, 2017, 2018, Nguyen and Mondelli, 2020].

Assumption 3.8 (Activation function)**.** *The activation function ϕ is non-linear and its derivative ϕ' is L-Lipschitz.*

We denote by μ_l' the l-th Hermite coefficient of ϕ'. We remark that the invertibility of the kernel K_{NTK} induced by the feature map (3.24) follows from Lemma D.13. At this point, we are ready to state our main result for the NTK model, whose full proof is contained in Appendix D.4.

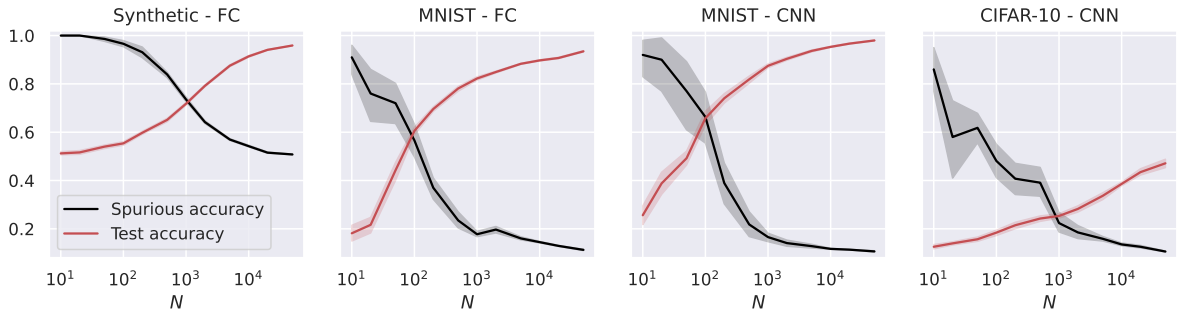


Figure 3.4: Test and spurious accuracies as a function of the number of training samples n, for a fully connected (FC, first two plots), and a small convolutional neural network (CNN, last two plots). In the first plot, we use synthetic (Gaussian) data with $d = 1000$, and the labeling function is $g(x) = \text{sign}(u^\top x)$. As we consider binary classification, the accuracy of random guessing is 0.5. The other plots use subsets of the MNIST and CIFAR-10 datasets, with an external layer of noise added to images, see Figure 3.1. As we consider 10 classes, the accuracy of random guessing is 0.1. We plot the average over 10 independent trials and the confidence band at 1 standard deviation.

Theorem 3.9. *Let Assumptions 3.3, 3.7, and 3.8 hold, and let $x \sim \mathcal{P}_X$ be sampled independently from everything. Consider querying the trained NTK model* (3.24) *with $z_1^s = [x, y_1]$. Let $\alpha = d_y/d \in (0,1)$ and $\mathcal{F}_{\text{NTK}}(z_1^s, z_1)$ be the feature alignment between z_1^s and z_1, as defined in* (3.8)*. Then, letting μ_l' be the l-th Hermite coefficient of the derivative of the activation ϕ' (see Appendix B), we have*

$$|\mathcal{F}_{\text{NTK}}(z_1^s, z_1) - \gamma_{\text{NTK}}| = o(1), \qquad \textit{where} \quad 0 < \gamma_{\text{NTK}} := \alpha \frac{\sum_{l=1}^{+\infty} \mu_l'^2 \alpha^l}{\sum_{l=1}^{+\infty} \mu_l'^2} < 1, \tag{3.26}$$

with probability at least $1 - n\exp(-c\log^2 k) - \exp(-c\log^2 n)$ over Z, x and W_0.

Theorem 3.6 proves the concentration of the feature alignment $\mathcal{F}_{\text{NTK}}(z_1^s, z_1)$ to a constant γ_{NTK}, which has an exact expression depending on α and the Hermite coefficients of the derivative of the activation. This result validates the argument in (3.10), where we replaced the feature alignment $\mathcal{F}_{\text{NTK}}(z_1^s, z_1)$ with a constant $\gamma_{\text{NTK}} > 0$. Thus, (3.11) now reads

$$\text{Cov}\left(f_{\text{NTK}}(z_1^s, \theta^*), g_1\right) \le \gamma_{\text{NTK}} \sqrt{\mathcal{R}_{Z_{-1}}} \sqrt{\text{Var}\left(g_1\right)}, \tag{3.27}$$

which quantifies the memorization of spurious features in terms of the generalization error and the constant γ_{NTK}.

Figure 3.3 considers training on MNIST and CIFAR-10, and it shows that the predictions of Theorem 3.9 also hold for standard datasets: as n increases, the test accuracy improves and the spurious accuracy decreases; considering activations with dominant high-order Hermite coefficients reduces memorization. For additional experiments using the same setup as Figure 3.2 and highlighting the dependence on α, we refer the reader to Figures D.1 and D.2 in Appendix D.5.

Proof sketch. The argument is more direct than for the RF model since, in this case, we are able to express γ_{NTK} in closed form. We denote by $P_{\Phi_{-1}}$ the projector over the span of

the rows of the NTK feature matrix $\Phi_{\mathsf{NTK},-1}$ without the first row. Then, the *first step* is to *center the feature map* φ_{NTK}, which gives

$$\frac{\varphi_{\mathsf{NTK}}(z_1^s)^\top P^\perp_{\Phi_{-1}} \varphi_{\mathsf{NTK}}(z_1)}{\left\| P^\perp_{\Phi_{-1}} \varphi_{\mathsf{NTK}}(z_1) \right\|_2^2} \simeq \frac{\tilde{\varphi}_{\mathsf{NTK}}(z_1^s)^\top P^\perp_{\Phi_{-1}} \tilde{\varphi}_{\mathsf{NTK}}(z_1)}{\left\| P^\perp_{\Phi_{-1}} \tilde{\varphi}_{\mathsf{NTK}}(z_1) \right\|_2^2}, \tag{3.28}$$

where $\tilde{\varphi}_{\mathsf{NTK}}(z) := z \otimes (\phi'(W_0 z) - \mathbb{E}_{W_0}[\phi'(W_0 z)])$. While a similar step appeared in the analysis of the RF model, its implementation for NTK requires a different strategy. In particular, we exploit that the samples z_1 and z_1^s are *approximately* contained in the span of the rows of Z_{-1} (see Lemma D.16). As the rows of Z_{-1} may not *exactly* span all $\mathbb{R}^d$, we resort to an *approximation* by adding a small amount of independent noise to every entry of Z_{-1}. The resulting perturbed dataset $\bar{Z}_{-1}$ satisfies $\mathrm{Span}\{\mathrm{rows}(\bar{Z}_{-1})\} = \mathbb{R}^d$ (see Lemma D.15), and we conclude via a continuity argument with respect to the magnitude of the perturbation (see Lemmas D.14 and D.17).

The *second step* is to upper bound the terms $\left|\tilde{\varphi}_{\mathsf{NTK}}(z_1^s)^\top P_{\Phi_{-1}} \tilde{\varphi}_{\mathsf{NTK}}(z_1)\right|$ and $\left\| P_{\Phi_{-1}} \tilde{\varphi}_{\mathsf{NTK}}(z_1) \right\|_2^2$, showing they have *negligible magnitude*, which gives

$$\frac{\tilde{\varphi}_{\mathsf{NTK}}(z_1^s)^\top P^\perp_{\Phi_{-1}} \tilde{\varphi}_{\mathsf{NTK}}(z_1)}{\left\| P^\perp_{\Phi_{-1}} \tilde{\varphi}_{\mathsf{NTK}}(z_1) \right\|_2^2} \simeq \frac{\tilde{\varphi}_{\mathsf{NTK}}(z_1^s)^\top \tilde{\varphi}_{\mathsf{NTK}}(z_1)}{\|\tilde{\varphi}_{\mathsf{NTK}}(z_1)\|_2^2}. \tag{3.29}$$

This is a consequence of the fact that, if $z \sim \mathcal{P}_Z$ is independent from Z_{-1}, then $\tilde{\varphi}(z)$ is roughly orthogonal to $\mathrm{Span}\{\mathrm{rows}(\Phi_{\mathsf{NTK},-1})\}$, see Lemma D.20.

Finally, the *third step* is to show the *concentration* over W_0 of the numerator and denominator of the RHS of (3.29), see Lemma D.18. This allows us to conclude that

$$\frac{\tilde{\varphi}_{\mathsf{NTK}}(z_1^s)^\top \tilde{\varphi}_{\mathsf{NTK}}(z_1)}{\|\tilde{\varphi}_{\mathsf{NTK}}(z_1)\|_2^2} \simeq \alpha \frac{\sum_{l=1}^{+\infty} \mu_l'^2 \alpha^i}{\sum_{l=1}^{+\infty} \mu_l'^2} > 0. \tag{3.30}$$

The RHS of (3.30) is strictly positive as $\alpha > 0$ and ϕ is non-linear by Assumption 3.8.

Discussion. As showed by (3.11) and (3.13), if $\mathcal{F}_\varphi(z_1, z_1^s)$ converges to 0, spurious correlations are *not* memorized by the model. However, Theorems 3.6 and 3.9 prove that, for the RF and NTK model respectively, a strictly positive geometric overlap ($\alpha > 0$) guarantees a strictly positive feature alignment. Thus, these results imply that, as long as the generalization error is not vanishing, spurious features are memorized and, in fact, we quantify the extent to which memorization occurs. Remarkably, our experiments on real datasets (see Figure 3.3) show that the effect of the activation function

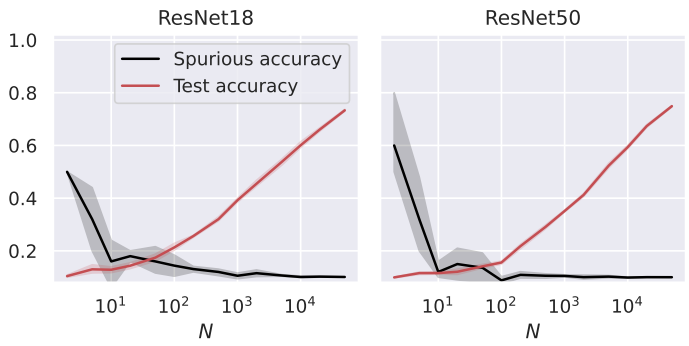


Figure 3.5: Test and spurious accuracies as a function of the number of training samples n, for two ResNet architectures. We use subsets of the CIFAR-10 dataset, with an external layer of noise added to images, see Figure 3.1. As we consider 10 classes, the accuracy of random guessing is 0.1. We plot the average over 10 independent trials and the confidence band at 1 standard deviation.

is in line with our predictions: taking ϕ' with higher order Hermite coefficients lead to models that are less prone to memorize spurious features. Finally, while we focus on generalized linear regression, the interplay between memorization of spurious features and generalization provided by our analysis holds in far more generality and, in particular, it is displayed by neural networks also capable of feature learning, see Figures 3.4 and 3.5.

3.7 Conclusions

In this chapter, we present a theoretical framework to quantify the memorization of spurious features. Our characterization hinges on *(i)* the classical notion of *stability* of the model w.r.t. a training sample, and *(ii)* a novel notion of *feature alignment* $\mathcal{F}(z^s, z)$ between two samples that share the same spurious feature y. By providing a precise analysis of the feature alignment in the prototypical settings of random and NTK features regression, we show that the memorization is proportional to the generalization error, and we characterize the proportionality constant, revealing how it depends on the model and its activation function. Our theoretical predictions are confirmed by numerical experiments on standard datasets (see Figure 3.3) and on different neural network architectures (see Figures 3.4 and 3.5).

The approach we put forward is rather general, and our results could be extended to cases where the spurious feature y is correlated with the ground-truth label g. Another possible extension involves testing the trained model on a new spurious feature y', which is not present in the training set but is correlated with a feature y that has already been seen; or capturing the role of *simplicity bias* in this phenomenon. We also remark that the formalism introduced by Lemma 3.2 applies to any feature map φ (*e.g.*, with multiple fully-connected, convolutional or attention layers). Characterizing the feature alignment of such maps would allow to compare different models and establish which of those is less prone to memorizing spurious correlations.

CHAPTER 4

A Law of Data Reconstruction for Random Features (and Beyond)

Large-scale deep learning models are known to *memorize* parts of the training set. In machine learning theory, memorization is often framed as interpolation or label fitting, and classical results show that this can be achieved when the number of parameters p in the model is larger than the number of training samples n. In this work, we consider memorization from the perspective of *data reconstruction*, demonstrating that this can be achieved when p is larger than dn, where d is the dimensionality of the data. More specifically, we show that, in the random features model, when $p \gg dn$, the subspace spanned by the training samples in feature space gives sufficient information to identify the individual samples in input space. Our analysis suggests an optimization method to reconstruct the dataset from the model parameters, and we demonstrate that this method performs well on various architectures (random features, two-layer fully-connected and deep residual networks). Our results reveal a *law of data reconstruction*, according to which the entire training dataset can be recovered as p exceeds the threshold dn.

This chapter is based on the work published at ICLR 2026: *A law of data reconstruction for random features (and beyond)*.

4.1 Introduction

How many parameters does a neural network need to memorize the training data?

The answer to this question depends on what one means by *memorization*, a term used with different purposes in the machine learning literature. Informally speaking, as described in the introduction, it captures the phenomenon of models storing the information of individual training samples, as opposed to learning the statistical patterns in the data. Thus, on the one hand, it is used to describe the phenomenon of label fitting [Zhang et al., 2017b] or forms of leave-one-out output stability [Feldman, 2020, Feldman and Zhang, 2020]. On the other hand, memorization is associated with *reconstructing* parts of the *training set* through knowledge of the model parameters [Carlini et al., 2023b, Schwarzschild et al., 2024, Cooper and Grimmelmann, 2024]. Notably, this reconstruction is possible in modern foundation models [Carlini et al., 2021, Nasr et al., 2025a, Cooper et al., 2025], with consequences in terms of

privacy concerns and copyright infringement [Tramèr et al., 2024, Cooper and Grimmelmann, 2024].

Understanding how many parameters a neural network needs to interpolate the dataset (or, equivalently, memorize the labels) is a classical problem [Cover, 1965], with more modern literature showing that interpolation occurs as soon as the number of model parameters p exceeds the number of training samples n [Soltanolkotabi et al., 2018, Montanari and Zhong, 2020, Bombari et al., 2022b]. An intuition for the phenomenon is that solving n equations (given by the training samples) generally requires $p \geq n$ degrees of freedom (given by the model parameters). In contrast, for the problem of reconstructing the training dataset, the situation is less clear. Empirical work has observed that this task becomes easier as the model gets larger [Haim et al., 2022, Carlini et al., 2023b]. However, theoretical work has taken different perspectives, focusing *e.g.* on impossibility results for differentially private training [Balle et al., 2022], on reconstructing the data from the model gradients [Wang et al., 2023] or on models with an infinite number of parameters [Loo et al., 2024]. To the best of our knowledge, there is no theoretical result connecting feasibility of data reconstruction and model size.

To address this gap, we propose a *law of data reconstruction*, giving a threshold for which reconstruction becomes possible. We consider *random features* (RF) regression, where the model is $f_{\mathsf{RF}}(x, \theta) = \varphi(x)^\top \theta$, see Eq. (4.1). Here, x is the d-dimensional input, $\varphi(x)$ the corresponding p-dimensional feature vector, and θ the p-dimensional model parameter vector obtained by running gradient descent on the square loss over n samples. This chapter gives theoretical and empirical evidence that:

$$\text{all } \textit{the training data can be reconstructed when } p \gg dn.$$

An intuition for the phenomenon is that solving dn equations (given by each dimension of each training sample) generally requires $p \geq dn$ degrees of freedom (given by the model parameters). Our analysis also suggests an optimization algorithm to reconstruct the training dataset and, as a proof of concept, we discuss the results of the designed method on CIFAR-10. We train a family of RF models using the square loss on $n = 100$ training samples. The left panel of Figure 4.1 shows the training loss in black and the *reconstruction error* in red (formally defined in Eq. (4.12)), as functions of the number of parameters p. As expected, the training error converges to 0 after the interpolation threshold $p = n$, where labels are memorized. In contrast, the reconstruction error drops only when p is of order dn. In the right panel of Figure 4.1, we display training images (odd rows) and reconstructed ones (even rows) for $p = 10dn$, showing that the whole dataset is recovered successfully. Our contributions are summarized below:

- In Theorem 4.4, we consider a set of n points $\hat{x}_1, \dots, \hat{x}_n \in \mathbb{R}^d$, and prove that when $p \gg dn$, if for every training sample x_i it holds $\varphi(x_i) \in \mathrm{Span}\{\{\varphi(\hat{x}_j)\}_{j=1}^n\}$, then all the $\hat{x}_j$-s *must be close* to one of the original training data.
- As the previous result does not exclude the possibility of duplicates within the $\hat{x}_j$-s, in Theorem 4.5 we prove that the $\hat{x}_j$-s *must be distinct*, focusing on the case $n = 2$ for simplicity. Taken together, Theorem 4.4 and 4.5 show that, when $p \gg dn$, the entire training set can be reconstructed given knowledge of the subspace $\mathrm{Span}\{\{\varphi(x_i)\}_{i=1}^n\}$. In fact, having access to $\mathrm{Span}\{\{\varphi(x_i)\}_{i=1}^n\}$, one can then look for $\hat{x}_1, \dots, \hat{x}_n$ s.t. $v \in \mathrm{Span}\{\{\varphi(\hat{x}_j)\}_{j=1}^n\}$ for any $v \in \mathrm{Span}\{\{\varphi(x_j)\}_{j=1}^n\}$.

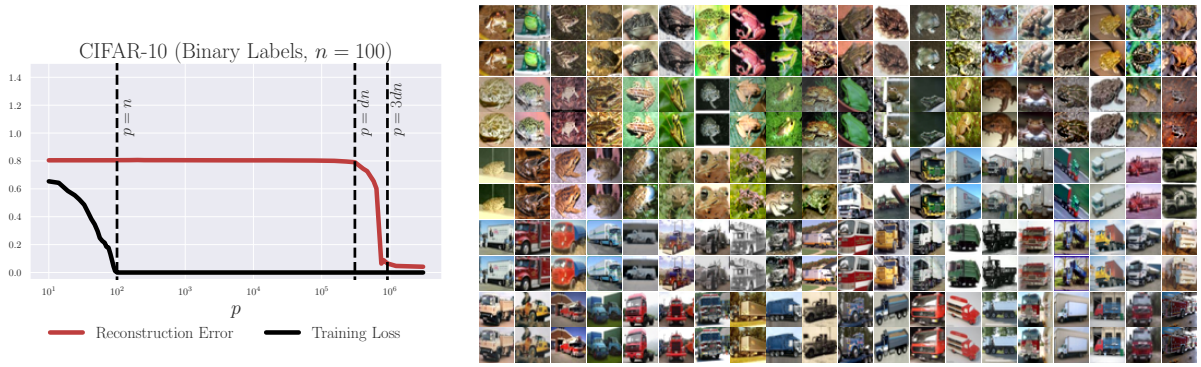


Figure 4.1: **Thresholds for label fitting and data reconstruction in the random features model.** (Left) We consider RF regression with ReLU activation on binary labels (*frogs vs. trucks*) with $n = 100$ images (50 examples per class) from CIFAR-10 ($d = 3072$). We report the mean for both the reconstruction error (in red, defined in (4.12)) and the training error (in black, mean squared error), as the number of parameters p increases. Statistics are computed across 4 distinct random seeds, and the standard deviation across seeds is very small (the confidence interval at one standard deviation is reported in the plot as shaded area, but it is imperceptible). For $p \geq n$, training labels are memorized, while reconstruction is feasible when p becomes larger than dn. (Right) Results of the reconstruction when $p = 10dn$. Odd rows report the ground truth images, while even rows the reconstructed ones which are all visually very similar.

- In practice, we have access to the vector of trained parameters θ^*, which is in $\mathrm{Span}\{\{\varphi(x_i)\}_{i=1}^n\}$, see Eq. (4.2). This motivates considering the *reconstruction loss* $\|P_{\hat{\Phi}}^{\perp}\theta^*\|_2^2$, where $P_{\hat{\Phi}}$ is the projector on $\mathrm{Span}\{\{\varphi(\hat{x}_j)\}_{i=1}^n\}$: if $\|P_{\hat{\Phi}}^{\perp}\theta^*\|_2^2 = 0$, then $\theta^* \in \mathrm{Span}\{\{\varphi(\hat{x}_j)\}_{j=1}^n\}$. We empirically show that optimizing this loss over the $\hat{x}_j$-s via gradient descent leads to the reconstruction of the training dataset, when $p \gg dn$. Notably, this procedure for data reconstruction is not limited to the RF model, but it performs well also for two-layer and deep residual networks, see Figures 4.6-4.7.

We finally remark that the scaling $p \gg dn$ was previously considered in the context of adversarial robustness: prior work proved that the condition $p \gg dn$ is both necessary [Bubeck et al., 2021, Bubeck and Sellke, 2021] and, in some settings, sufficient [Bombari et al., 2023] for *smooth* label interpolation. This suggests an inherent connection between the adversarial robustness of the model and the ability to reconstruct training data from knowledge of its parameters. Our code is publicly available at `https://github.com/iurada/data-reconstruction-law`.

4.2 Related work

Over-parameterization and memorization. The problem of label memorization and how it relates to the number of parameters in a neural network dates back to seminal results [Cover, 1965, Baum, 1988]. More recently, a line of work aimed at showing that gradient descent converges to 0 training loss (i.e., the model memorizes the labels) for networks with progressively smaller over-parameterization, as discussed in Chapter 2. There, we show that $p \gg n$ parameters are sufficient to fit any set of labels in deep neural networks. Under a separability assumption, it has also been shown that n arbitrary training points can be perfectly classified by a neural network with $p \gg \sqrt{n}$ parameters [Vardi et al., 2022]. This, in the context of regression, implies that $p \gg \log(1/\epsilon)\sqrt{n}$ is sufficient to guarantee an error of at most ϵ. In contrast, the problem of input data memorization and reconstruction has been much less explored from a theoretical standpoint. Balle et al. [2022] proved impossibility

results if the model is trained with differential privacy [Dwork et al., 2006], and showed that in generalized linear models it is possible to reconstruct an individual data provided all other training samples are known. Loo et al. [2024] considered networks with an infinite number of parameters, Smorodinsky et al. [2024] focused on partial uni-dimensional ($d = 1$) training set reconstruction, and Wang et al. [2023] assumed knowledge of model gradients, requiring $p \gg d^2$ parameters with n at most of order $d^{1/4}$. The regime $p \gg dn$ was heuristically identified by Haim et al. [2022] as enabling successful data reconstruction, while [Brown et al., 2021, Feldman et al., 2025] studied its role as a requirement for learning certain tasks, rather than for data reconstruction.

Data reconstruction. Recent studies have shown that training data can be extracted from generative models by carefully prompting their generation routines [Carlini et al., 2023a, Zhang et al., 2023, Nasr et al., 2025a, Cooper et al., 2025]. More generally, data reconstruction is a broader task, not limited to generative models, that aims to recover the training set directly from the learned model's parameters [Cooper and Grimmelmann, 2024]. A possible approach is via model inversion attacks, which optimize over the input space to raise a class score, although this method generally recovers class prototypes rather than specific training samples [Fredrikson et al., 2015, Mahendran and Vedaldi, 2015]. Another strategy relies on the implicit bias of gradient descent for homogeneous networks [Lyu and Li, 2020, Ji and Telgarsky, 2020]: Haim et al. [2022] proposed a reconstruction objective motivated by the observation that gradient descent converges to points satisfying the KKT conditions of a max-margin problem, and empirically showed that training samples on the classification margin can be recovered. This method was then extended by Buzaglo et al. [2023], Oz et al. [2024] to the multiclass setting with general losses, and to larger scale pretrained models. Loo et al. [2024] adapted the idea to neural networks trained with the square loss in the linear (or NTK) regime [Jacot et al., 2018, Lee et al., 2019]. This leads to an optimization method for data reconstruction which has similarities with ours (compare Eq. (7) in [Loo et al., 2024] with Eq. (4.11) in this chapter).

We refer the reader to Chapter 3 for a discussion of the related literature on the RF model.

4.3 Problem setup

Notation. All complexity notations $\omega, \Omega, \Theta, O, o$ are understood for sufficiently large input dimension d and number of parameters p. Given a positive number k, $[k]$ denotes the set of positive numbers from 1 to k. Given a vector v, $\|v\|_2$ denotes its Euclidean norm. Given a matrix $A \in \mathbb{R}^{m\times n}$, we denote by $P_A \in \mathbb{R}^{n\times n}$ the projector over $\mathrm{Span}\{\mathrm{rows}(A)\}$. We say that an event holds with overwhelming probability if it holds with probability at least $1 - e^{-\omega(\log d)}$.

Let (X, Y) be a labeled training dataset, where $X = [x_1, \dots, x_n]^\top \in \mathbb{R}^{n\times d}$ contains the training input data (sampled i.i.d. from a distribution $\mathcal{P}_X$) on its rows and $Y = (y_1, \dots, y_n) \in \mathbb{R}^n$ contains the corresponding labels. We define the *random features* (RF) model as

$$f_{\mathsf{RF}}(x, \theta) = \phi(Vx)^\top \theta = \varphi(x)^\top \theta, \tag{4.1}$$

where $\varphi(x) \in \mathbb{R}^p$ is the feature vector associated to the input $x \in \mathbb{R}^d$, and $\theta \in \mathbb{R}^p$ are trainable parameters. The random features matrix $V \in \mathbb{R}^{p\times d}$ is s.t. $V_{i,j} \sim_{\text{i.i.d.}} \mathcal{N}(0, 1/d)$, and $\phi : \mathbb{R} \to \mathbb{R}$ is a non-linearity applied component-wise. We note that $f_{\mathsf{RF}}(x, \theta)$ is a generalized linear model, and it can be regarded as a two-layer fully-connected neural network, where only the second layer is trained. We consider the supervised learning setup where a quadratic

training loss (without regularization) is minimized via gradient descent. When the learning rate is sufficiently small, gradient descent converges to the interpolator which is the closest in ℓ_2 norm to the initialization (see Equation (33) in [Bartlett et al., 2021]), which we consider equal to 0 for simplicity. Then, denoting with $\Phi := [\varphi(x_1), \dots, \varphi(x_n)]^\top \in \mathbb{R}^{n \times p}$ the feature matrix, the trained parameters are

$$\theta^* = \Phi^+ Y, \tag{4.2}$$

where Φ^+ is the Moore-Penrose inverse of Φ.

Assumption 4.1 (Data distribution)**.** *The training samples $\{x_1, \dots, x_n\}$ are n i.i.d. samples from the sub-Gaussian distribution $\mathcal{P}_X$, with $\|x\|_{\psi_2} = O(1)$, and such that $\|x\|_2 = \sqrt{d}$.*

This assumption requires the data to have well-behaved tails (see the definition of the sub-Gaussian norm $\|\cdot\|_{\psi_2}$ in Eq. (B.1) in Appendix B for further details), and allows for badly conditioned distributions. This hypothesis is commonly used in the related literature, and it includes *e.g.* Gaussian data, data respecting Lipschitz concentration [Bubeck and Sellke, 2021, Bombari et al., 2023, von Berg et al., 2025], and data uniform on the sphere [Mei and Montanari, 2022, Hu et al., 2024]. The normalization $\|x\|_2 = \sqrt{d}$ is chosen for technical convenience, and this scaling of the norm (combined with the scaling of the random features matrix V) guarantees that the pre-activations of the model (*i.e.* , the entries of Vx) are of constant order.

Assumption 4.2 (Activation function)**.** *The activation function $\phi : \mathbb{R} \to \mathbb{R}$ is a non-linear, Lipschitz continuous function such that its derivative is also Lipschitz. Letting μ_l denote the l-th Hermite coefficient of ϕ, we further assume that $\mu_0 = \mu_2 = 0$, $\mu_1 \neq 0$, and that there exist two non-zero Hermite coefficients of order ≥ 3 with different parity.*

This assumption is motivated by theoretical convenience, and we expect our results to hold for a wider set of activations (as in [Mei et al., 2021]) after a more involved analysis. Except for the last condition on the parity of high-order Hermite coefficients, Assumption 4.2 resembles the setting considered by Hu and Lu [2022], and it covers a wide family of odd activations, including $\tanh$. The parity of high-order Hermite coefficients is used to reconstruct the correct sign of the training data, and it appears to be necessary for that, see Remark 4.4 for details.

Assumption 4.3 (Over-parameterization)**.** *We consider a data dimensionality regime where $n = O(d)$, and an over-parameterized model with*

$$p = \omega\left(nd \log^2 d\right). \tag{4.3}$$

To guarantee that the RF model interpolates the data, it suffices that $p \gg n$ [Mei et al., 2021, Wang and Zhu, 2021]. The stronger over-parameterization requirement in Eq. (4.3) was shown to be both necessary [Bubeck and Sellke, 2021] and, in some settings, sufficient [Bombari et al., 2023] to achieve *smooth* interpolation. We focus on the regime $n = O(d)$, which includes the popular proportional regime $n = \Theta(d)$ as well as the regime where n is a fixed constant (independent of d, p). We expect our results to be generalizable also to $n \gg d$ [Mei et al., 2021, Hu et al., 2024, Pandit et al., 2024].

4.4 Main results

In this section, we present our main theoretical results, pinpointing $p \gg dn$ as the over-parameterization threshold such that data reconstruction can take place given knowledge of the subspace of the training samples in feature space.

More formally, the goal of data reconstruction is to exhibit a matrix $\hat{X} \in \mathbb{R}^{n \times d}$ such that its rows $\hat{x}_j \in \mathbb{R}^d$ are close to the rows of the original training data matrix X. When $\|\hat{x}_j\|_2 = \|x_i\|_2 = \sqrt{d}$, a reconstructed sample $\hat{x}_j$ can be considered close to a training sample x_i if $\|\hat{x}_j - x_i\|_2 = o(\sqrt{d})$. Let us also define $\hat{\Phi} = [\varphi(\hat{x}_1), \dots, \varphi(\hat{x}_n)] \in \mathbb{R}^{n \times p}$ as the feature matrix of the reconstructed data. Then, the result below gives sufficient conditions for all rows of $\hat{X}$ to be close to training samples.

Theorem 4.4. *Let Assumptions 4.1, 4.2, and 4.3 hold. Let $\hat{X} \in \mathbb{R}^{n \times d}$ be such that its rows satisfy $\|\hat{x}_i\|_2 = \sqrt{d}$, and for every $i \in [n]$, $\varphi(x_i) \in Span\{\mathrm{rows}(\hat{\Phi})\}$. Then, with overwhelming probability, for any $\hat{\imath} \in [n]$, there exists $i \in [n]$ such that*

$$\|\hat{x}_{\hat{\imath}} - x_i\|_2 = o(\sqrt{d}). \tag{4.4}$$

In words, Theorem 4.4 states that, if the random features of the training samples are spanned by the random features of a matrix $\hat{X}$, the rows of $\hat{X}$ *must* be close (in input space) to the original training samples. Geometrically, this result proves that, when $p \gg dn$, in order to span the subspace generated by the training data features, one has to consider approximately the same vectors, as there is no solution to this problem obtained by non-degenerate linear combinations. In contrast with [Loo et al., 2024] which tackles the case $p \to \infty$, our main contribution is to pinpoint the threshold $p \gg dn$ constituting a sufficient amount of over-parameterization to reconstruct the data. We also highlight that, for the claim of Theorem 4.4 to hold, assumptions on the activation are necessary: if ϕ is linear, picking $\hat{x}_i = \sqrt{d}(x_i + x_{i+1})/\|x_i + x_{i+1}\|_2$ ($i \in [n-1]$) and $\hat{x}_n = \sqrt{d}(x_n - x_1)/\|x_n - x_1\|_2$ gives a counterexample to (4.4). The proof of Theorem 4.4 is deferred to Appendix E.1, and a sketch is below.

Proof sketch. Prior results on the RF kernel concentration [Mei et al., 2021, Wang and Zhu, 2021] guarantee that, for $p \gg n$, the smallest eigenvalue $\lambda_{\min}(\Phi\Phi^\top)$ is bounded away from 0. Thus, the rows of Φ are linearly independent (they span a sub-space of dimension *exactly* n). Then, as the n row vectors of $\hat{\Phi}$ span all rows of Φ (by hypothesis), $\mathrm{Span}\{\mathrm{rows}(\Phi)\} = \mathrm{Span}\{\mathrm{rows}(\hat{\Phi})\}$. This in turn gives that, if $\hat{x} \in \mathbb{R}^d$ is a generic row of $\hat{X}$, then $\varphi(\hat{x}) \in \mathrm{Span}\{\mathrm{rows}(\Phi)\}$, i.e.,

$$\varphi(\hat{x}) = \sum_{i=1}^{n} a_i \varphi(x_i). \tag{4.5}$$

As argued above, the result cannot hold for linear activations, so let us focus on the non-linear component $\tilde{\varphi}(x_i) = \phi(Vx_i) - \mu_1 Vx_i$ in the Hermite basis. Taking the inner product of both sides of (4.5) with $\tilde{\varphi}(x_i)$ and with $\tilde{\varphi}(\hat{x})$ yields, with overwhelming probability,

$$\left|\tilde{\varphi}(x_i)^\top \tilde{\varphi}(\hat{x}) - \tilde{\mu}^2 a_i\right| = \tilde{O}\left(\sqrt{\frac{dn}{p}}\right) + o(1), \qquad \left|\|a\|_2^2 - 1\right| = \tilde{O}\left(\sqrt{\frac{dn}{p}}\right) + o(1), \tag{4.6}$$

where $a \in \mathbb{R}^n$ is defined as the vector containing a_i in its i-th entry, and $\tilde{\mu}^2$ is the sum of the squares of the Hermite coefficients of ϕ of order at least 3. The two results in Eq. (4.6) are formalized in Lemmas E.2-E.3, and they rely on concentration arguments on the random features V, which have to hold *uniformly* over any $\hat{x} \in \sqrt{d}\,\mathbb{S}^{d-1}$. To obtain such uniform concentration, we rely on an ε-net argument which crucially uses the condition $p \gg dn$.

Next, in Lemma E.4, we show that, with overwhelming probability,

$$\text{if} \quad C = \max_i \left|\frac{x_i^\top \hat{x}}{d}\right|, \qquad \text{then} \quad \left\|\frac{(X\hat{x})^{\circ l}}{d^l}\right\|_2 \le C^{l-1} + o(1), \tag{4.7}$$

uniformly for every $l \geq 2$. Combining (4.6)-(4.7) gives that $|C-1| = o(1)$, i.e., there exists a single training sample x_j aligned with $\hat{x}$. To resolve the ambiguity in the sign, we use that there exist two non-zero Hermite coefficients of order ≥ 3 with different parity, which concludes the argument. □

Sign ambiguity. The existence of two non-zero Hermite coefficients with different parity is a necessary condition to recover the sign of the training samples. In fact, if ϕ is either even or odd, the problem is under-determined in terms of the sign of the $\hat{x}_i$-s, as $\mathrm{Span}\{\mathrm{rows}(\hat{\Phi})\}$ does not depend on them. Remarkably, this effect is also evident in numerical experiments optimizing the reconstruction loss defined in Eq. (4.11): Figure 4.4 considers ReLU activation (which violates the last condition of Assumption 4.2, as its Hermite coefficients $\mu_{2l+1} = 0$ for all $l > 1$) showing that negatives of training samples may be reconstructed.

Theorem 4.4 guarantees that all the rows of $\hat{X}$ are close to the training samples. However, it may still happen that multiple rows of $\hat{X}$ are close to the same sample, leaving part of the training dataset not reconstructed. This gap is approached by our next result which focuses on the case $n = 2$.

Theorem 4.5. *Let Assumptions 4.1, 4.2, and 4.3 hold. Let $n = 2$ and $\hat{X} \in \mathbb{R}^{2\times d}$ be such that its rows satisfy $\|\hat{x}_i\|_2 = \sqrt{d}$, and for every $i \in \{1,2\}$, $\varphi(x_i) \in Span\{\mathrm{rows}(\hat{\Phi})\}$. Then, with overwhelming probability, for any $i \in \{1,2\}$, there exists $\hat{\imath} \in \{1,2\}$ such that*

$$\|\hat{x}_{\hat{\imath}} - x_i\|_2 = o(\sqrt{d}). \tag{4.8}$$

In words, Theorem 4.5 shows that *all* training samples are in fact reconstructed, ruling out the possibility of $\hat{X}$ containing repetitions. The proof is deferred to Appendix E.2 and a sketch is below.

Proof sketch. We approach the problem by contradiction, supposing the existence of two vectors $\varepsilon_1, \varepsilon_2 \in \mathbb{R}^d$ such that

$$\|\varepsilon_1\|_2, \|\varepsilon_2\|_2 = o(\sqrt{d}), \qquad \varphi(x_2) = a_1\varphi(x_1+\varepsilon_1) + a_2\varphi(x_1+\varepsilon_2),$$

for some real values a_1 and a_2. A Taylor expansion of $\varphi(x_1+\varepsilon_2)$ around $x_1+\varepsilon_1$ gives

$$\begin{aligned}\varphi(x_2) = (a_1+a_2)\varphi(x_1+\varepsilon_1) \\ + a_2\,\phi'(V(x_1+\varepsilon_1)) \circ (V(\varepsilon_2-\varepsilon_1)) \\ + a_2\left((\varphi(x_1+\varepsilon_2)-\varphi(x_1+\varepsilon_1)) - \phi'(V(x_1+\varepsilon_1)) \circ (V(\varepsilon_2-\varepsilon_1))\right),\end{aligned} \tag{4.9}$$

where $\circ$ denotes the component-wise product of two vectors. First, we take the inner product of both sides of Eq. (4.9) with Vx_1, which yields

$$|a_1+a_2| = O\left(\frac{|a_2|\,\|\varepsilon_2-\varepsilon_1\|_2}{\sqrt{d}}\right) + o(1),$$

with overwhelming probability (see Lemma E.9). Then, we take the inner product with $V(\varepsilon_2-\varepsilon_1)$, and show that the first and third term of the RHS of Eq. (4.9) are negligible with respect to the second one (see Lemmas E.7 and E.8). An upper bound on the LHS of Eq. (4.9) based on Cauchy-Schwartz inequality is then enough to show in Lemma E.10 that

$$|a_2| = O\left(\frac{\sqrt{d}}{\|\varepsilon_2-\varepsilon_1\|_2}\right), \qquad |a_1+a_2| = O(1),$$

with overwhelming probability. The argument in Lemma E.7 relies on a concentration result on the sum of independent random variables with sub-exponential norm much larger than their standard deviation (see Lemma E.6). This is obtained via a Bernstein-type bound combined with an ε-net argument as, similarly to Theorem 4.4, we need a *uniform* control over any choice of $\varepsilon_1, \varepsilon_2$. Finally, by taking the inner product of the two sides of Eq. (4.9) with $\tilde{\varphi}(x_2)$, we obtain

$$\tilde{\varphi}(x_2)^\top \varphi(x_2) \leq \left|(a_2 + a_2)\tilde{\varphi}(x_2)^\top \varphi(x_1 + \varepsilon_1)\right| + \left|a_2 \tilde{\varphi}(x_2)^\top \left(\varphi(x_1 + \varepsilon_2) - \varphi(x_1 + \varepsilon_1)\right)\right| . \tag{4.10}$$

Now, the LHS of Eq. (4.10) is $\Theta(p)$ since ϕ is non-linear; the first term in the RHS of Eq. (4.10) is $o(p)$ as x_1 and x_2 are roughly orthogonal with overwhelming probability; and the last term in the RHS of Eq. (4.10) is also $o(p)$ via generalized Stein's lemma (see Lemma E.11). This gives the desired contradiction. □

Technical challenge for $n \geq 3$. We note that, already when $n = 3$, the presence of duplicates is either given by the "triplet" $\hat{x}_1, \hat{x}_2, \hat{x}_3$ all similar to each other, or by a pair of duplicates $\hat{x}_1, \hat{x}_2$, with a different $\hat{x}_3$. The constructive approach used in the proof of Theorem 4.5 would require us to consider the two cases separately, and the amount of cases increases combinatorially with n. Nevertheless, we suspect the idea that the span of duplicate samples cannot contain higher order terms of the left-out samples ($\tilde{\varphi}(x_2)$ in Eq. (4.10)) to carry over to a general n, as long as $p \gg dn$.

From the theory to a reconstruction algorithm. Our theoretical analysis shows that, under sufficient over-parameterization ($p \gg dn$), the matrix $\hat{X}$ successfully reconstructs the training dataset when $\|P_{\hat{\Phi}}^{\perp} \varphi(x_i)\|_2 = 0$ for every $i \in [n]$, where $P_{\hat{\Phi}}$ denotes the projector on $\mathrm{Span}\{\mathrm{rows}(\hat{\Phi})\}$. In practice, we only have access to the trained model θ^*, V and the activation ϕ. Recall $\theta^* = \Phi^+ Y \in \mathrm{Span}\{\mathrm{rows}(\Phi)\}$, so θ^* is a linear combination of $\{\varphi(x_i)\}_{i=1}^n$, which suggests to solve the problem:

$$\hat{X}^* = \underset{\hat{X} \,:\, \|\hat{x}_i\|_2 = \sqrt{d}}{\arg\min} \mathcal{L}(\hat{X}) \ , \quad \mathcal{L}(\hat{X}) = \left\| P_{\hat{\Phi}}^{\perp} \theta^* \right\|_2^2 . \tag{4.11}$$

Importantly, enforcing that θ^* lies in the span of the reconstructed features ($\mathcal{L}(\hat{X}) = 0$) does not immediately imply that $\varphi(x_i) \in \mathrm{Span}\{\mathrm{rows}(\hat{\Phi})\}$ for all i (as the implication only goes in the other direction). In Figure 4.2, we numerically minimize $\mathcal{L}(\hat{X})$ and check whether the vectors $\varphi(x_i)$ approximately lie in the subspace $\mathrm{Span}\{\mathrm{rows}(\hat{\Phi})\}$. To do so, we calculate the average per-feature orthogonal residual, *i.e.* the average over $i \in [n]$ of $\|P_{\hat{\Phi}}^{\perp} \varphi(x_i)\|_2 / \sqrt{p}$, where $P_{\hat{\Phi}}^{\perp} = I - P_{\hat{\Phi}}$ projects onto the orthogonal complement of $\mathrm{Span}\{\mathrm{rows}(\hat{\Phi})\}$. This quantity equals 0 if and only if every $\varphi(x_i)$ lies in $\mathrm{Span}\{\mathrm{rows}(\hat{\Phi})\}$. The normalization by $\sqrt{p}$ makes $r(\hat{\Phi}) := \sum_{i=1}^n \|P_{\hat{\Phi}}^{\perp} \varphi(x_i)\|_2 / (n\sqrt{p})$ of order 1 so that values

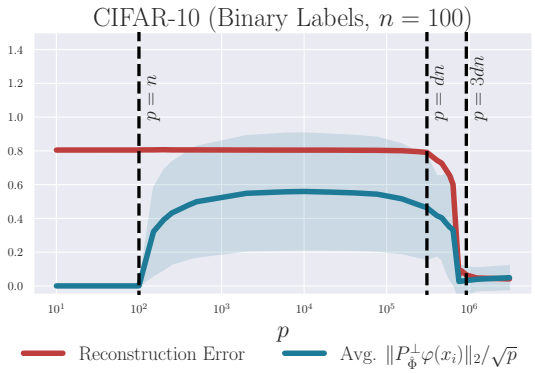


Figure 4.2: **Features of the training dataset Φ are spanned by the features of the reconstructed dataset $\hat{\Phi}$.** We consider the same setup as in Figure 4.1. For different values of p, we optimize until $\mathcal{L}(\hat{X}) = 0$, and report reconstruction error (in red, defined in Eq. (4.12)) and normalized residual $\|P_{\hat{\Phi}}^{\perp} \varphi(x_i)\|_2$ averaged over $i \in [n]$ (in blue), with their confidence interval at one standard deviation (shaded area). Further details and evidence are in Appendix E.4.1, see Figure E.1.

of $r(\hat{\Phi}) \ll 1$ indicate numerically negligible residuals (*i.e.* , effective span inclusion). In Figure 4.2, the per-feature orthogonal residual (plotted in blue) remains large until the model crosses the threshold $p \approx dn$, after which it drops sharply (for $p \leq n$, the optimization converges to the non-degenerate case $P_{\hat{\Phi}} = I$, which makes the orthogonal residual trivially zero). Thus, this numerical evidence suggests that minimizing $\mathcal{L}(\hat{X})$ is sufficient to satisfy the hypotheses of our main theorems for $p \gg dn$, and hence to successfully reconstruct the training data. Equipped with these insights and a recipe for dataset reconstruction, we complement our theory with numerical results as discussed below.

4.5 Numerical experiments

We initialize the rows $\hat{x}_i$ of $\hat{X}$ with i.i.d. standard Gaussian vectors. We then minimize $\mathcal{L}(\hat{X})$ in (4.11) with gradient descent with momentum, normalizing each row $\hat{x}_i$ after the update to constrain it on the d-dimensional sphere $\sqrt{d}\,\mathbb{S}^{d-1}$. In every experiment, we perform the optimization until the reconstruction loss converges to zero, *i.e.* , $\mathcal{L}(\hat{X}^*) = 0$ (up to machine precision). To quantify reconstruction quality, we report the average ℓ_2 distance between the rows of the ground truth and reconstructed data matrices, modulo a permutation on the rows of the latter, *i.e.* ,

$$\rho(X, \hat{X}^*) = \min_{\Pi \in \mathcal{P}_n} \frac{1}{n\sqrt{d}} \sum_{i=1}^{n} \left\| x_i - \hat{x}^*_{\Pi(i)} \right\|_2 , \tag{4.12}$$

where $\mathcal{P}_n$ denotes the set of permutations of $[n]$. Obtaining Π^* constitutes a classic linear assignment problem [Bertsekas, 1998], which we solve in polynomial time via the Hungarian method [Kuhn, 1955]. We normalize our success metric with $n\sqrt{d}$ so that a reconstruction can be considered successful as $\rho(X, \hat{X}^*)$ becomes much smaller than 1. As our main focus is on ReLU activations, whose odd Hermite coefficients (except μ_1) are zero, recovered samples can appear with flipped sign (see Remark 4.4 and Figure 4.4). Accordingly, we also maximize $\rho(X, \hat{X}^*)$ over per-image sign flips.

In the following, we present numerical results by first considering a synthetic setup (i.i.d. data uniformly distributed on the d-dimensional sphere), and then by reconstructing natural images from the CIFAR-10 dataset (Section 4.5.1). Finally, we explore the extent to which these phenomena are observed in neural networks trained with gradient descent (Section 4.5.2).

4.5.1 Random features regression

Synthetic data on the d-dimensional sphere. We begin with a synthetic task, where the training data $X = [x_1, \dots, x_n]^\top \in \mathbb{R}^{n \times d}$ are n i.i.d. samples drawn uniformly from the sphere of radius $\sqrt{d}$. The labels $Y \in \mathbb{R}^n$ are given by $Y = Xg + \epsilon$, where $g \in \mathbb{R}^d$ has entries $g_i \sim \mathcal{N}(0, 1/d)$ and the noise $\epsilon \in \mathbb{R}^n$ is independent of the data X, with i.i.d. Gaussian entries with zero mean and variance 0.25. Figure 4.3 showcases the same trend

Figure 4.5: **Reconstruction of higher-dimensional images.** We repeat the same experiment of Figure 4.1 (right) using $n = 20$ samples from Tiny-ImageNet and a random features model with $p = 10dn$.

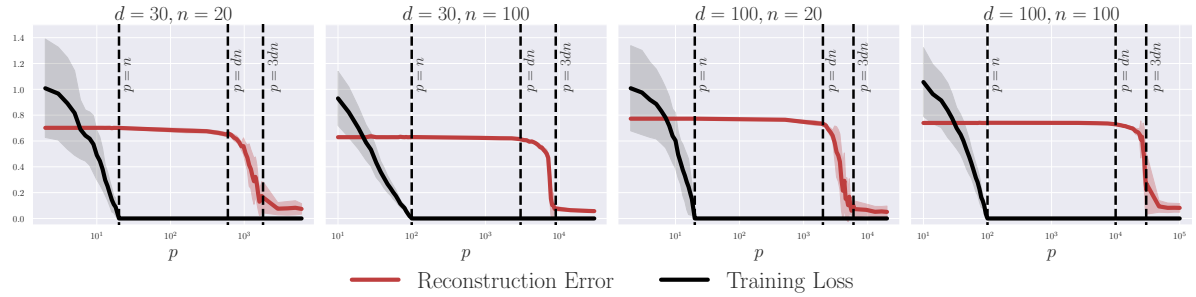


Figure 4.3: **Thresholds for label fitting and data reconstruction when training on i.i.d. data unformly drawn from the d-dimensional sphere.** We consider RF regression with ReLU activation, fitting a noisy linear model. We report mean (solid line) and standard deviation (shaded area) for both the reconstruction error (in red) and training loss (in black) as the number of parameters p increases, at different choices of input dimensions d and number of dataset examples n. Statistics are computed across 10 distinct random seeds. Two distinct thresholds clearly emerge: $p \gg n$ for label fitting, and $p \gg dn$ for data reconstruction.

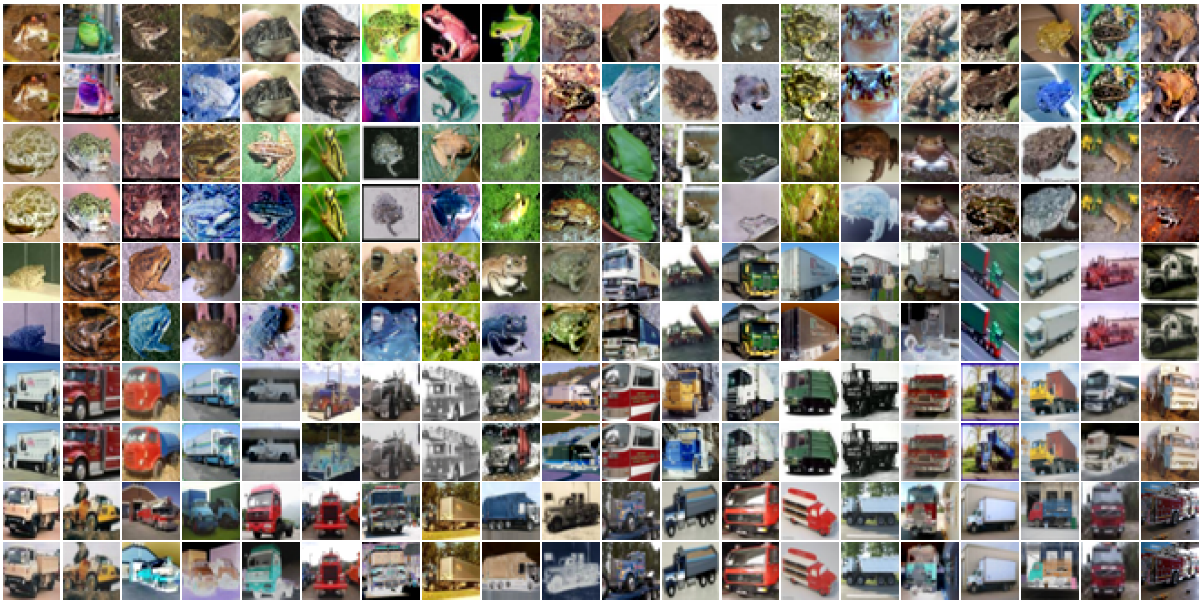


Figure 4.4: **Images reconstructed from an RF model with ReLU activation may have the wrong sign.** We repeat the experiment of Figure 4.1 using a different random seed and observe that reconstructions from ReLU models can appear as sign-flipped versions of original training data. This is due to the fact that ReLU has odd Hermite coefficients of order ≥ 3 equal to zero, as discussed in Remark 4.4.

for several choices of d and n: at $p \geq n$, the training loss approaches zero, while the reconstruction error approaches zero when p becomes larger than dn.

Natural images with binary labels. The same phenomenon observed in the previous scenario carries over to natural images with binary labels $\{\pm 1\}$. For these experiments, we restrict the CIFAR-10 training split to the first 50 instances of the "*frog*" and "*truck*" classes, fitting several RF models with ReLU activation. In the left part of Figure 4.1, we can appreciate the same transitions at $p = n$ and $p \approx dn$ respectively for label fitting and data reconstruction; the right part of the figure then demonstrates that the reconstructed dataset appears perceptually indistinguishable from the original training dataset. Notably, for ReLU the above visual match may fail due to a *sign error*. In fact, with a different seed, we reconstruct the negatives

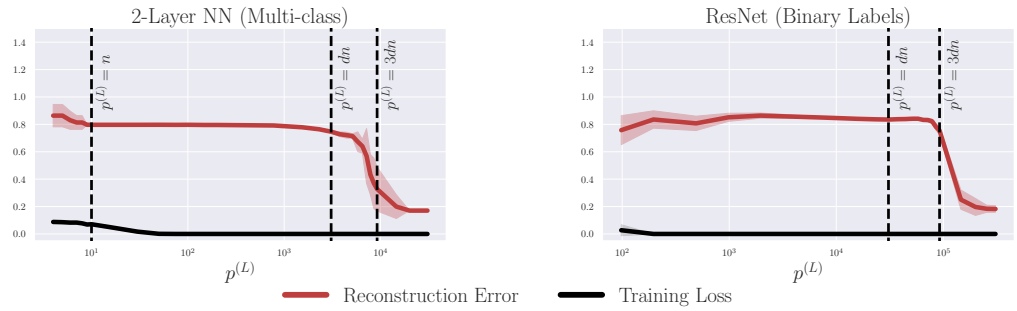


Figure 4.6: **Thresholds for label fitting and data reconstruction for neural networks trained with gradient descent on $n = 10$ CIFAR-10 images.** We consider regression with the square loss, training two-layer ReLU networks on the one-hot encoding of the 10-class labels (left) and ResNets on binary labels (right). We report mean (solid line) and standard deviation (shaded area) for both reconstruction error (in red) and training loss (in black), as the number of parameters in the last layer $p^{(L)}$ increases. Statistics are computed across 10 distinct random seeds.

of some training images (Figure 4.4) even if the condition $p \gg dn$ is met and the loss in Eq. (4.11) converges. This aligns with Remark 4.4: ReLU violates Assumption 4.2 since $\mu_{2\ell+1} = 0$ for $\ell > 1$. In Figure E.2 deferred to Appendix E.4.2, we show that the sign ambiguity disappears upon taking the activation $\phi(z) = \mathrm{ReLU}(z) + \tanh(z)$, which has mixed-parity Hermite coefficients and, therefore, satisfies Assumption 4.2. In Figure 4.5, we display the reconstructed images from the same experiment on Tiny-ImageNet, which has input dimension $d = 3 \times 64 \times 64$, for $n = 20$ (additional details can be found in Figure E.6 and Appendices E.3–E.4.3).

4.5.2 Neural networks trained with gradient descent

Motivated by the RF baseline, we now turn our attention to finite-width networks, studying how *the number of parameters in the last layer* $p^{(L)}$ determines the ability to reconstruct the training dataset. Concretely, we train two-layer and deep residual networks with full-batch gradient descent, minimizing the square loss. Given that in this case the initialization is non-zero, we modify the reconstruction loss as $\mathcal{L}(\hat{X}) = \|P^{\perp}_{\hat{\Phi}}(\theta^{(L)}_* - \theta^{(L)}_0)\|_2^2$, where $\theta^{(L)}_0$ and $\theta^{(L)}_*$ are respectively the parameters of the last layer at initialization and at the end of training. The feature matrix $\hat{\Phi}$ stacks the penultimate layer feature vectors $\varphi(\hat{x}_i)$. We assume access to trained weights of all layers $\{\theta^{(1)}_*, \dots, \theta^{(L)}_*\}$, the last layer at initialization $\theta^{(L)}_0$, and the neural network's computational graph.

Two-layer neural networks. We consider a two-layer neural network $f_{\mathsf{NN}}(x) = \theta^{(2)}\phi(\theta^{(1)}x)$, with k-class output and ReLU activation ϕ. The weight matrices $\theta^{(1)} \in \mathbb{R}^{h\times d}$ and $\theta^{(2)} \in \mathbb{R}^{k\times h}$ have i.i.d. entries, initialized as $\theta^{(1)}_{i,j} \sim \mathcal{N}(0, 1/d)$, $\theta^{(2)}_{i,j} \sim \mathcal{N}(0, 1/h)$. In this setting, $p^{(L)} = k \times h$, where h is the width of the network. In the left panel of Figure 4.6, we analyze the ability to reconstruct CIFAR-10 images with one-hot targets for each of the 10 classes. Although the output is no longer a scalar (as for random features), when $p^{(L)}$ becomes larger than dn, the reconstruction error starts decreasing. For all models reported in the plot, the total number of trainable parameters satisfies $p > n$ (even when $p^{(L)} < n$), so the model interpolates all training labels and the training loss is close to zero. As a confirmation of the quality of reconstructed images, we report in Figure 4.7 the reconstruction of $n = 100$ examples (10 per class), when $p^{(L)} = 4dn$. Also in this scenario, the reconstructed images are perceptually indistinguishable from original training data.

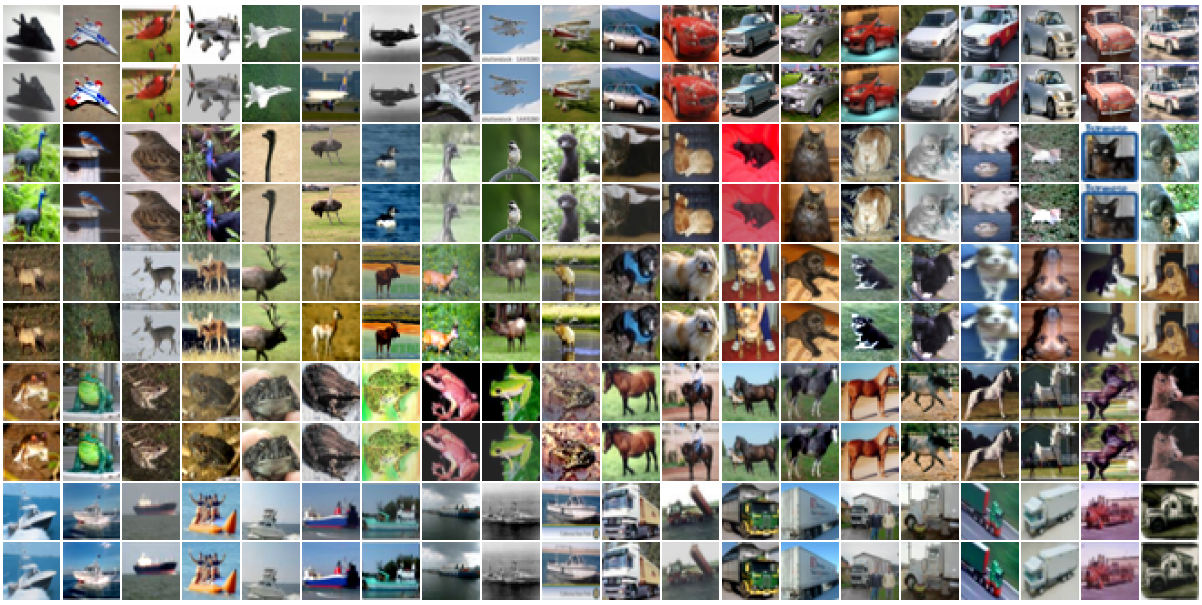


Figure 4.7: **Multi-class training data reconstructed from a neural network trained on CIFAR-10.** We train a two-layer ReLU network on $n = 100$ images from CIFAR-10 dataset (10 examples per class) with gradient descent. We cast the training procedure as multi-class regression on square loss, using one-hot encoded class labels as targets. The number of parameters in the last layer of the network is $p^{(L)} = 4dn$.

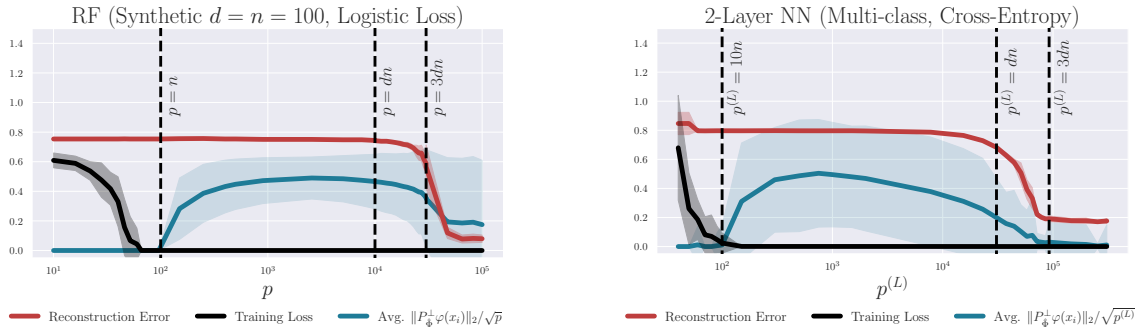


Figure 4.8: **Thresholds for data reconstruction on classification experiments.** (Left) We consider binary classification over $n = 100$ i.i.d. samples uniformly drawn from the d-dimensional sphere ($d = 100$). We train RF models with ReLU activation by minimizing the logistic loss with gradient descent. (Right) We consider multi-class classification and train two-layer ReLU networks on $n = 10$ CIFAR-10 images by minimizing cross-entropy loss with gradient descent. We report mean (solid line) and standard deviation (shaded area) for the reconstruction error (in red), for the training loss (in black) and for the per-feature orthogonal residual of projecting Φ onto $\mathrm{Span}\{\mathrm{rows}(\hat{\Phi})\}$ (in blue). We indicate with $p^{(L)}$ the number of parameters in the last layer. Statistics are computed across 10 distinct random seeds.

Deep residual networks. We conclude the experimental evaluation with residual architectures, probing how their structure (*i.e.* , residual connections and convolutions) affects reconstructability. We defer the formal definition of the model involved in these experiments to Appendix E.3.1. On CIFAR-10 (*frogs vs. trucks*), shown in the rightmost panel of Figure 4.6, ResNets display a transition for data reconstruction consistent with earlier experiments and happening after the $p^{(L)} = dn$ threshold. As for two-layer networks, the ResNets interpolate training labels even when $p^{(L)} < n$, since the total number of parameters p is much larger than the number of samples n for all models.

Classification via logistic and cross-entropy loss. We consider a synthetic task, where the training data x_i are $n = 100$ i.i.d. samples drawn uniformly on the sphere of radius $\sqrt{d} = 10$. The labels are given by $y_i = \text{sign}(g^\top x_i)$, where $g \in \mathbb{R}^d$ has entries $g_i \sim \mathcal{N}(0, 1/d)$. We compute θ^* minimizing the logistic loss $\ell(\theta) := \sum_{i=1}^n \log(1 + e^{-y_i \varphi(x_i)^\top \theta})$ with gradient descent, and consider the same reconstruction algorithm as in Eq. (4.11). In the left panel of Figure 4.8, we find the usual threshold $p \approx dn$ for successful reconstruction of the training set. In the right panel of Figure 4.8, we consider a two-layer neural network trained with cross-entropy loss on $n = 10$ samples from the 10 classes of CIFAR-10. and we see that the reconstruction algorithm yields similar results as the ones for regression in Figure 4.6. Further implementation details can be found in Appendix E.4.3, and the reconstructed images can be found in Figure E.9.

Additional ablation studies. In Appendix E.4.3, we first show that the optimization of $\mathcal{L}(\hat{X})$ is robust to different choices of the learning rate η (if it is sufficiently small). Then, we explore the setting where $\hat{X}$ has $\hat{n} \neq n$ rows, showing that the result of the optimization gives overlapping images in the case $\hat{n} < n$, and successfully reconstructs the full training set (plus some extra duplicates) in the case $\hat{n} > n$. We later show successful reconstruction from the parameters of a vision transformer, and give empirical evidence that adding a weight decay regularizer does not affect the law of reconstruction. Finally, we show that it is possible to reconstruct the data from a pruned neural network, albeit with higher values of p depending on the sparsity.

4.6 Conclusions

This chapter studies data memorization, intended as the feasibility of reconstructing n data samples of dimension d from a trained model, focusing on the number of parameters p at which this becomes possible. Our results on random features point to a *law of data reconstruction*, establishing the threshold $p \approx dn$. Remarkably, the reconstruction method grounded in our theoretical analysis is also successful in two-layer and deep residual networks. While our analysis assumes knowledge of the subspace of the training samples in feature space ($\text{Span}\{\{\varphi(x_i)\}_{i=1}^n\}$), the experimental results demonstrate that, by optimizing the loss $\mathcal{L}(\hat{X})$ in Eq. (4.11), the whole training set is reconstructed when $p \gg dn$. This suggests two outstanding open problems to be tackled in future research: proving that *(i)* all the global optima of $\mathcal{L}(\hat{X})$ are permutations of the training dataset, and that *(ii)* the optimization problem can be efficiently solved with gradient methods despite the non-convexity of $\mathcal{L}(\hat{X})$. Another interesting direction is to understand what happens in the regime $n \ll p \ll dn$. On the one hand, Balle et al. [2022] indicate how to reconstruct *any single* training sample in this regime under certain model assumptions, given access to the final parameters and, additionally, to the remaining training data. On the other hand, we cautiously suspect that, without this additional knowledge, reconstructing the entire dataset is information-theoretically impossible when $p \ll dn$: at fixed machine precision, this would require recovering $\Theta(dn)$ bits from only $\Theta(p)$ bits, underscoring a hard limit on reconstructability. Importantly, this would not guarantee that *none* of the training samples is memorized. There is empirical evidence that learning models tend to memorize outliers [Feldman, 2020], and a possible direction for future work is to investigate if parts of the training data sampled from a heavy-tailed distribution (therefore not respecting Assumption 4.1) would result in memorization before the threshold $p \approx dn$.

CHAPTER 5

Beyond the Universal Law of Robustness: Sharper Laws for Random Features and Neural Tangent Kernels

Machine learning models are vulnerable to adversarial perturbations, and a thought-provoking paper by Bubeck and Sellke has analyzed this phenomenon through the lens of over-parameterization: interpolating smoothly the data requires significantly more parameters than simply memorizing it. However, this "universal" law provides only a necessary condition for robustness, and it is unable to discriminate between models. In this chapter, we address these gaps by focusing on empirical risk minimization in two prototypical settings, namely, random features and the neural tangent kernel (NTK). We prove that, for random features, the model is not robust for any degree of over-parameterization, even when the necessary condition coming from the universal law of robustness is satisfied. In contrast, for even activations, the NTK model meets the universal lower bound, and it is robust as soon as the necessary condition on over-parameterization is fulfilled. This also addresses a conjecture in prior work by Bubeck, Li and Nagaraj. Our analysis decouples the effect of the kernel of the model from an "interaction matrix", which describes the interaction with the test data and captures the effect of the activation. Our theoretical results are corroborated by numerical evidence on both synthetic and standard datasets (MNIST, CIFAR-10).

This chapter is based on the work published at ICML 2024: *Beyond the universal law of robustness: Sharper laws for random features and neural tangent kernels*.

5.1 Introduction

Despite the deployment of deep neural networks in a wide set of applications, these models are known to be vulnerable to adversarial perturbations [Biggio et al., 2013, Szegedy et al., 2014], which raises serious concerns about their robustness guarantees. To address these issues, a wide variety of adversarial training methods has been developed [Goodfellow et al., 2015, Kurakin et al., 2017, Madry et al., 2018, Raghunathan et al., 2018, Wong and Kolter, 2018]. In a parallel effort, a recent line of theoretical work has focused on providing a principled understanding to the phenomenon of adversarial robustness, see e.g. Dohmatob [2022], Zhu et al. [2022], Javanmard et al. [2020b], Donhauser et al. [2021] and the review in Section

5.2. Specifically, the recent paper by [Bubeck and Sellke, 2021] has highlighted the role of over-parameterization as a *necessary* condition to achieve robustness: while interpolation requires the number of parameters p of the model to be at least linear in the number of data samples n, *smooth* interpolation (namely, robustness) requires $p > dn$, where d is the input dimension. We note that the law of robustness put forward by [Bubeck and Sellke, 2021] is "universal" in the sense that it holds for *any* parametric model. Furthermore, an earlier conjecture by [Bubeck et al., 2021] posits that there exists a model of two-layer neural network such that robustness is successfully achieved with this minimal number of parameters; i.e., the condition $p > dn$ is both *necessary* and *sufficient* for robustness. This state of affairs leads to the following natural question:

When does over-parameterization become a sufficient condition
to achieve adversarial robustness?

In this chapter, we consider a *sensitivity*[1] measure proportional to $\|\nabla_z f(z, \theta)\|_2$, where z is the test data point and f is the model with parameters θ whose robustness is investigated. This quantity is closely related to notions of sensitivity appearing in related works [Dohmatob, 2022, Dohmatob and Bietti, 2022, Bubeck and Sellke, 2021], and it leads to a more stringent requirement on the robustness than the perturbation stability considered by [Zhu et al., 2022], see also the discussion in Section 5.3. For such a sensitivity measure, we show that the answer to the question above depends on the specific model at hand. Specifically, we will focus on the solution yielded by empirical risk minimization (ERM) in two prototypical settings widely analyzed in the theoretical literature: *(i)* Random Features (RF) [Rahimi and Recht, 2007], and the *(ii)* Neural Tangent Kernel (NTK) [Jacot et al., 2018].

Main contribution. Our key results are summarized below.

- For random features, we show that ERM leads to a model which is *not robust* for *any* degree of over-parameterization. Specifically, we tackle a regime in which the universal law of robustness by [Bubeck and Sellke, 2021] trivializes, and we provide a more refined bound.
- For NTK with an even activation function, we give an upper bound on the ERM solution that *matches* the lower bound by [Bubeck and Sellke, 2021]. As the NTK model approaches the behavior of gradient descent for a suitable initialization [Chizat et al., 2019], this also shows that a class of two-layer neural networks has sensitivity of order $\sqrt{nd/p}$. This addresses Conjecture 2 of [Bubeck et al., 2021], albeit for a slightly different notion of sensitivity (see the comparison in Section 5.3).

At the technical level, our analysis involves the study of the spectra of RF and NTK random matrices, and it provides the following insights.

- We introduce in (5.19) a new quantity, dubbed the *interaction matrix*. Studying this matrix in isolation is a key step in all our results, as its different norms are intimately related to the sensitivity of both the RF and the NTK model. To the best of our

[1] The sensitivity can be interpreted as the non-robustness.

knowledge, this is the first time that attention is raised over such an object, which we deem relevant to the theoretical characterization of adversarial robustness.

- Our analytical characterization captures the role of the activation function: it turns out that a specific symmetry (e.g., being even) boosts the adversarial robustness of the models.

Finally, we experimentally show that the robustness behavior of both synthetic and standard datasets (MNIST, CIFAR-10) agrees well with our theoretical results. Our code is available at the GitHub repository `simone-bombari/beyond-universal-robustness`.

5.2 Related work

From the vast literature studying adversarial examples, we succintly review related works focusing on linear models, random features and NTK models.

Linear regression. In light of the universal law of robustness, in the interpolation regime, linear regression cannot achieve adversarial robustness (as $p = d$). For linear models, Donhauser et al. [2021], Javanmard et al. [2020b] have focused on adversarial training (as opposed to ERM, which is considered in this work), showing that over-parameterization hurts the robust generalization error. We also point out that, in some settings, even the standard generalization error is maximized when the model is under-parametrized [Hastie et al., 2022]. Precise asymptotics on the robust error in the classification setting are provided in Javanmard and Soltanolkotabi [2022], Taheri et al. [2020]. A different approach is pursued by Tsipras et al. [2019], who study a linear model that exhibits a trade-off between generalization and robustness.

Random features (RF). Beyond the more general discussion on the related literature on this model in Chapter 3, theoretical results on the adversarial robustness of random features have been shown for both adversarial training [Hassani and Javanmard, 2024] and the ERM solution [Dohmatob, 2022, Dohmatob and Bietti, 2022]. We will discuss the comparison with Dohmatob [2022], Dohmatob and Bietti [2022] in the next paragraph.

Neural tangent kernel (NTK). Beyond the more general discussion on the related literature on this object in Chapter 2, the NTK has also been recently exploited to identify non-robust features learned with standard training [Tsilivis and Kempe, 2022] and gain insight on adversarial training [Loo et al., 2022]. More closely related to this chapter, the robustness of the NTK model is studied in Zhu et al. [2022], Dohmatob and Bietti [2022], Dohmatob [2022]. Specifically, Zhu et al. [2022] analyze the interplay between width, depth and initialization of the network, thus improving upon results by Huang et al. [2021b], Wu et al. [2021]. However, the perturbation stability considered by Zhu et al. [2022] captures an average robustness, rather than an adversarial one (see the detailed comparison in Section 5.3), which makes these results not immediately comparable to ours. In contrast, our notion of sensitivity is close to that analyzed in Dohmatob and Bietti [2022], Dohmatob [2022], which establish trade-offs between robustness and memorization for both the RF and the NTK model. However, Dohmatob and Bietti [2022] study a teacher-student model with quadratic teacher and infinite data, which also does not allow for a direct comparison. Finally, Dohmatob [2022] studies the ERM solution and focuses on *(i)* infinite-width networks, and *(ii)* finite-width networks where the number of neurons scales linearly with the input dimension. We remark that this last setup

does not lead to lower bounds on the sensitivity tighter than those by Bubeck and Sellke [2021], and our results on the NTK model provide the first upper bounds on the sensitivity (which also match the lower bounds by Bubeck and Sellke [2021]).

5.3 Preliminaries

Notation. Given a vector v, we indicate with $\|v\|_2$ its Euclidean norm. Given $v \in \mathbb{R}^{d_v}$ and $u \in \mathbb{R}^{d_u}$, we denote by $v \otimes u \in \mathbb{R}^{d_v d_u}$ their Kronecker product. Given a matrix A, let $\|A\|_{\text{op}}$ be its operator norm, $\|A\|_F$ its Frobenius norm and $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ its smallest and largest eigenvalues respectively. We denote by $\|\varphi\|_{\text{Lip}}$ the Lipschitz constant of the function φ. All the complexity notations $\Omega(\cdot)$, $\mathcal{O}(\cdot)$, $o(\cdot)$ and $\Theta(\cdot)$ are understood for sufficiently large data size n, input dimension d, number of neurons k, and number of parameters p. We indicate with $C, c > 0$ numerical constants, independent of n, d, k, p.

Generalized linear regression. Let (X, Y) be a labelled training dataset, where $X = [x_1, \dots, x_n]^\top \in \mathbb{R}^{n \times d}$ contains the training samples on its rows and $Y = (y_1, \dots, y_n) \in \mathbb{R}^n$ contains the corresponding labels. Let $\Phi : \mathbb{R}^d \to \mathbb{R}^p$ be a generic *feature map*, from the input space to a feature space of dimension p. We consider the following *generalized linear model*

$$f(x, \theta) = \Phi(x)^\top \theta, \tag{5.1}$$

where $\Phi(x) \in \mathbb{R}^p$ is the feature vector associated with the input sample x, and $\theta \in \mathbb{R}^p$ is the set of the trainable parameters of the model. Our supervised learning setting involves solving the optimization problem

$$\min_\theta \left\| \Phi(X)^\top \theta - Y \right\|_2^2, \tag{5.2}$$

where $\Phi(X) \in \mathbb{R}^{n \times p}$ is the feature matrix, containing $\Phi(x_i)$ in its i-th row. From now on, we use the shorthands $\Phi := \Phi(X)$ and $K := \Phi\Phi^\top \in \mathbb{R}^{n \times n}$, where K denotes the kernel associated with the feature map. If we assume K to be invertible (i.e., the model is able to fit any set of labels Y), it is well known that gradient descent converges to the interpolator which is the closest in ℓ_2 norm to the initialization, see e.g. [Gunasekar et al., 2017]. In formulas,

$$\theta^* = \theta_0 + \Phi^\top K^{-1}(Y - f(X, \theta_0)), \tag{5.3}$$

where θ^* is the gradient descent solution, θ_0 is the initialization and $f(X, \theta_0) = \Phi(X)^\top \theta_0$ is the output of the model (5.1) at initialization.

Sensitivity. We measure the robustness of a model $f(z, \theta)$ as a function of the test sample z. In particular, we are interested in a quantity that expresses how *sensitive* is the output when small perturbations are applied to the input z. More specifically, we could imagine an adversarial example "built around" z as

$$z_{\text{adv}} = z + \Delta_{\text{adv}}, \quad \text{with } \|\Delta_{\text{adv}}\|_2 \le \delta \|z\|_2 . \tag{5.4}$$

Here, Δ_{adv} is a *small* adversarial perturbation, in the sense that its ℓ_2 norm is only a fraction δ of the ℓ_2 norm of z. Crucially, we expect δ not to depend on the scalings of the problem. For example, if z represents an image with d pixels, this can correspond to perturbing every pixel by a given amount. In this case, the robustness depends on the amount of perturbation per pixel, and not on the number of pixels d. This in turn implies that δ is a numerical constant independent of d.

We are interested in the response of the model to z_{adv}, and how this relates to the natural scaling of the output, which we assume to be $\Theta(1)$.[2] Up to first order, we can write

$$\begin{aligned} |f(z_{\text{adv}}, \theta) - f(z, \theta)| &\simeq \left|\nabla_z f(z,\theta)^\top \Delta_{\text{adv}}\right| \\ &\leq \|\Delta_{\text{adv}}\| \, \|\nabla_z f(z,\theta)\|_2 \\ &\leq \delta \, \|z\|_2 \, \|\nabla_z f(z,\theta)\|_2 . \end{aligned} \tag{5.5}$$

The first inequality is saturated if the attacker builds an adversarial perturbation that perfectly aligns with $\nabla_z f(z,\theta)$, and the second inequality follows from (5.4). Hence, we define the *sensitivity* of the model evaluated in z as

$$\mathcal{S}_{f_\theta}(z) := \|z\|_2 \, \|\nabla_z f(z,\theta)\|_2 . \tag{5.6}$$

Recall that both δ and the output of the model are constants, independent of the scaling of the problem. Hence, a robust model needs to respect $\mathcal{S}_{f_\theta}(z) = O(1)$ for its test samples z. In contrast, if $\mathcal{S}_{f_\theta}(z) \gg 1$ (e.g., the sensitivity grows with the number of pixels of the image), we expect the model to be adversarially vulnerable when evaluated in z. In a nutshell, the goal of this chapter is to provide bounds on $\mathcal{S}_{f_\theta}(z)$ for random features and NTK models, thus establishing their robustness.

Related notions of sensitivity. Measures of robustness similar to (5.6) have been used in the related literature. In particular, [Dohmatob and Bietti, 2022] study $\mathbb{E}_{z\sim P_X}[\mathcal{S}^2_{f_\theta}(z)]$ in the context of a teacher-student setting with a quadratic target. [Dohmatob, 2022] considers the model obtained from solving the generalized regression problem (5.2) and characterizes a sensitivity measure given by the Sobolev semi-norm. This quantity is again similar to $\mathbb{E}_{z\sim P_X}[\mathcal{S}^2_{f_\theta}(z)]$, the only difference being that the gradient is projected onto the sphere (namely, onto the data manifold) before taking the norm. [Zhu et al., 2022] study a notion of perturbation stability given by $\mathbb{E}_{z,\hat{z},\theta}\left|\nabla_z f(z,\theta)^\top (z - \hat{z})\right|$, where z is sampled from the data distribution and $\hat{z}$ is uniform in a ball centered at z with radius δ. This would be similar to averaging over Δ_{adv} in (5.5), instead of choosing it adversarially. Hence, [Zhu et al., 2022] capture a weaker notion of *average robustness*, as opposed to this chapter which studies the *adversarial robustness*. Finally, [Bubeck et al., 2021, Bubeck and Sellke, 2021] consider $\|f(z,\theta)\|_{\text{Lip}}$, where the Lipschitz constant is with respect to z, which is equivalent to $\max_z \mathcal{S}_{f_\theta}(z)/\|z\|_2$. While this is a more stringent condition than (5.6), we remark that the adversarial robustness corresponds to choosing adversarially Δ_{adv} (and not z), as done in (5.5). We also note that our main results will hold with high probability over the distribution of z.

To conclude, we remark that, differently from previous work [Bubeck and Sellke, 2021, Dohmatob, 2022], we opt for a *scale invariant* definition of sensitivity. This derives from the term $\|z\|_2$ appearing in the RHS of (5.6), and it allows us to apply the definition to data with arbitrary scaling. In particular, if we set $\|z\|_2 = 1$, then (5.4) would yield $\|\Delta_{\text{adv}}\|_2 \leq \delta$. Both [Bubeck and Sellke, 2021] and [Dohmatob, 2022] consider this scaling of the data: the former paper uses $O(1)$ Lipschitz constant as an indication of a robust model, and the latter considers a term similar to $\|\nabla_z f(z,\theta)\|_2$.

Robustness of generalized linear regression. For generalized linear models with feature map Φ, plugging (5.1) in (5.6) gives

$$\mathcal{S}_{f_\theta}(z) = \|z\|_2 \left\|\nabla_z \Phi(z)^\top \theta\right\|_2 . \tag{5.7}$$

[2] This is a natural choice, as e.g. the output label is not expected to grow with the input dimension. However, any scaling is in principle allowed and would not change our final results.

This expression can also provide the sensitivity of the model trained with gradient descent. Let us assume that $f(x, \theta_0) = 0$ for all $x \in \mathbb{R}^d$ and, as before, that the kernel of the feature map is invertible. Then, by plugging the value of θ^* from (5.3) in the previous equation, we get

$$\mathcal{S}_{f_{\theta^*}}(z) = \|z\|_2 \left\| \nabla_z \Phi(z)^\top \Phi^\top K^{-1} Y \right\|_2 . \tag{5.8}$$

5.4 Main Result for Random Features

In this section, we provide our law of robustness for the *random features* model. In particular, we will show that, for a wide class of activations, the sensitivity of random features is $\gg 1$ and, therefore, the *model is not robust*.

We assume the data labels $Y \in \mathbb{R}^n$ to be given by $Y = g(X) + \epsilon$. Here, $g : \mathbb{R}^d \to \mathbb{R}$ is the ground-truth function applied row-wise to $X \in \mathbb{R}^{n \times d}$ and $\epsilon \in \mathbb{R}^n$ is a noise vector independent of X, having independent sub-Gaussian entries with zero mean and variance $\mathbb{E}[\epsilon_i^2] = \varepsilon^2$ for all i. The *random features (RF) model* takes the form

$$f_{\mathsf{RF}}(x, \theta) = \Phi_{\mathsf{RF}}(x)^\top \theta, \qquad \Phi_{\mathsf{RF}}(x) = \phi(Vx), \tag{5.9}$$

where V is a $k \times d$ matrix, such that $V_{i,j} \sim_{\text{i.i.d.}} \mathcal{N}(0, 1/d)$, and ϕ is an activation function applied component-wise. The number of parameters of this model is k, as V is fixed and $\theta \in \mathbb{R}^k$ contains the trainable parameters. We initialize the model at $\theta_0 = 0$ so that $f_{\mathsf{RF}}(x, \theta_0) = 0$ for all x. After training, we can write the sensitivity using (5.8), which gives

$$\mathcal{S}_{\mathsf{RF}}(z) = \|z\|_2 \left\| \nabla_z \Phi_{\mathsf{RF}}(z)^\top \Phi_{\mathsf{RF}}^\top K_{\mathsf{RF}}^{-1} Y \right\|_2 , \tag{5.10}$$

where we use the shorthands $\Phi_{\mathsf{RF}} := \Phi_{\mathsf{RF}}(X) = \phi(XV^\top) \in \mathbb{R}^{n \times k}$ for the RF map and $K_{\mathsf{RF}} := \Phi_{\mathsf{RF}} \Phi_{\mathsf{RF}}^\top \in \mathbb{R}^{n \times n}$ for the RF kernel. We will show that the kernel is in fact invertible in the argument of our main result, Theorem 5.5. Throughout this section, we make the following assumptions.

Assumption 5.1 (Data distribution)**.** *The input data $(x_1, \dots, x_n)$ are n i.i.d. samples from the distribution P_X, which satisfies the following properties:*

1. $\|x\|_2 = \sqrt{d}$*, i.e., the data are distributed on the sphere of radius* $\sqrt{d}$*.*
2. $\mathbb{E}[x] = 0$*, i.e., the data are centered.*
3. P_X *satisfies the* Lipschitz concentration property*. Namely, there exists an absolute constant $c > 0$ such that, for every Lipschitz continuous function $\varphi : \mathbb{R}^d \to \mathbb{R}$, we havefor all $t > 0$,*
$$\mathbb{P}\left(|\varphi(x) - \mathbb{E}_X[\varphi(x)]| > t\right) \le 2e^{-ct^2/\|\varphi\|_{\text{Lip}}^2}. \tag{5.11}$$

The first two assumptions can be achieved by simply pre-processing the raw data. For the third assumption, we remark that the family of Lipschitz concentrated distributions covers a number of important cases, e.g., standard Gaussian [Vershynin, 2018], uniform on the sphere and on the unit (binary or continuous) hypercube [Vershynin, 2018], or data obtained via a Generative Adversarial Network (GAN) [Seddik et al., 2020]. The third requirement is common in the related literature [Nguyen et al., 2021b, Bombari et al., 2022b] and it is equivalent to the isoperimetry condition required by Bubeck and Sellke [2021]. We remark that the three

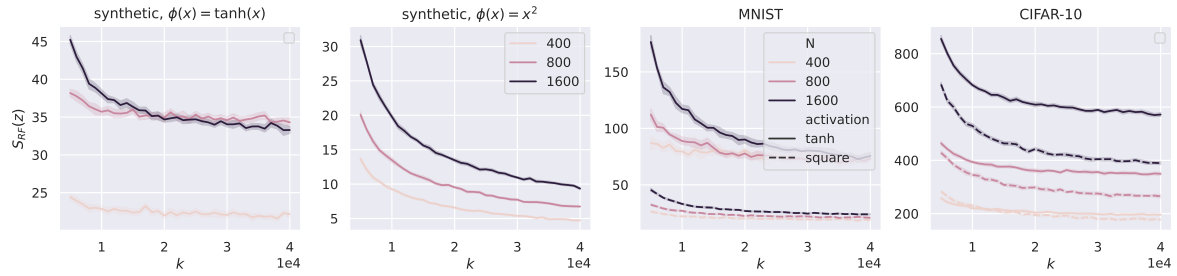


Figure 5.1: Sensitivity as a function of the number of parameters k, for an RF model with synthetic data sampled from a Gaussian distribution with input dimension $d = 1000$ (first and second plot), and with inputs taken from two classes of the MNIST and CIFAR-10 datasets (third and fourth plot respectively). Different curves correspond to different values of the sample size $n \in \{400, 800, 1600\}$ and to different activations ($\phi(x) = \tanh(x)$ or $\phi(x) = x^2$). We plot the average over 10 independent trials and the confidence interval at 1 standard deviation. The values of the sensitivity are significantly larger for $\phi(x) = \tanh(x)$ than for $\phi(x) = x^2$.

conditions of Assumption 5.1 are often replaced by a stronger requirement (e.g., data uniform on the sphere), see Montanari and Zhong [2020], Dohmatob [2022].

Finally, we note that our choice of data scaling is different from the related work [Bubeck and Sellke, 2021, Dohmatob, 2022], as they both consider $\|z\|_2 = 1$. We choose $\|z\|_2 = \sqrt{d}$, as we believe it to be more natural in practice. Consider, for example, the case in which z is an image and d is the number of pixels. If the pixel values are renormalized in a fixed range, then $\|z\|_2$ scales as $\sqrt{d}$. This choice shouldn't worry the reader, as our definition of sensitivity (5.6) is scale invariant, facilitating the comparison between previous results and ours.

Assumption 5.2 (Activation function)**.** *The activation function ϕ satisfies the following properties:*

1. *ϕ is a non-linear L-Lipschitz function.*
2. *Its first and second order derivatives ϕ' and ϕ'' are respectively L_1 and L_2-Lipschitz functions.*

These requirements are satisfied by common activations, e.g. smoothed ReLU, sigmoid, or $\tanh$.

Assumption 5.3 (Over-parameterization)**.**

$$n \log^3 n = o(k), \tag{5.12}$$
$$d \log^4 d = o(k). \tag{5.13}$$

In words, the number of neurons k scales faster than the input dimension d and the number of data points n. Such an over-parameterized setting allows to *(i)* perfectly interpolate the data (this follows directly from the invertibility of the kernel, which can be readily deduced from the proof of Theorem 5.5), and *(ii)* achieve minimum test error, see Mei and Montanari [2022]. This is also in line with the current trend of heavily over-parametrized deep-learning models [Nakkiran et al., 2020].

Assumption 5.4 (High-dimensional data)**.**

$$n \log^3 n = o\left(d^{3/2}\right), \tag{5.14}$$

$$\log^8 k = o(d). \tag{5.15}$$

While the second condition is rather mild, the first one lower bounds the input dimension d in a way that appears to be crucial to prove our main Theorem 5.5. Understanding whether (5.14) can be relaxed is left as an open question.

At this point, we are ready to state our main result for random features.

Theorem 5.5. *Let Assumptions 5.1, 5.2, 5.3 and 5.4 hold, and let $Y = g(X) + \epsilon$, such that the entries of ϵ are sub-Gaussian with zero mean and variance $\mathbb{E}[\epsilon_i^2] = \varepsilon^2$ for all i, independent between each other and from everything else. Let $z \sim P_X$ be a test sample independent from the training set (X, Y), and let the derivative of the activation function satisfy $\mathbb{E}_{\rho\sim\mathcal{N}(0,1)}[\phi'(\rho)] \neq 0$. Define $\mathcal{S}_{\mathsf{RF}}(z)$ as in* (5.10)*. Then,*

$$\mathcal{S}_{\mathsf{RF}}(z) = \Omega\left(\sqrt[6]{n}\right) \gg 1, \tag{5.16}$$

with probability at least $1 - \exp(-c \log^2 n)$ over X, z, V and ϵ.

In words, Theorem 5.5 shows that a vast class of over-parameterized RF models is not robust against adversarial perturbations. A few remarks are now in order.

Beyond the universal law of robustness. We recall that the results by Bubeck and Sellke [2021] give a lower bound on the sensitivity of order $\sqrt{nd/k}$. A lower bound of the same order is obtained by Dohmatob [2022], whose results in the finite-width setting require n, d, k to follow a proportional scaling (i.e., $n = \Theta(d) = \Theta(k)$). This leaves as an open question the characterization of the robustness for sufficiently over-parameterized random features (i.e., when $k \gg nd$). Our Theorem 5.5 resolves this question by showing that the RF model is in fact not robust for *any* degree of over-parameterization (as long as $k \gg n, d$ and Assumption 5.4 is satisfied).

Impact of the activation function and numerical results. For the result of Theorem 5.5 to hold, we require the additional assumption $\mathbb{E}_{\rho\sim\mathcal{N}(0,1)}[\phi'(\rho)] \neq 0$. This may not be just a technical requirement. In fact, if $\mathbb{E}_{\rho\sim\mathcal{N}(0,1)}[\phi'(\rho)] = 0$, then a key term appearing in the lower bound for the sensitivity (i.e., the Frobenius norm of the *interaction matrix* defined in (5.19)) has a drastically different scaling, see Theorem 5.8 in the proof outline below. The importance of the condition $\mathbb{E}_{\rho\sim\mathcal{N}(0,1)}[\phi'(\rho)] \neq 0$ is also confirmed by our numerical simulations in the synthetic setting considered in Figure 5.1. The first plot (corresponding to $\phi(x) = \tanh(x)$ s.t. $\mathbb{E}_{\rho\sim\mathcal{N}(0,1)}[\phi'(\rho)] \neq 0$) displays values of the sensitivity which are significantly larger than those in the second plot (corresponding to $\phi(x) = x^2$ s.t. $\mathbb{E}_{\rho\sim\mathcal{N}(0,1)}[\phi'(\rho)] = 0$). Furthermore, the sensitivity for $\phi(x) = \tanh(x)$ appears to reach a plateau for large values of k, while it keeps decreasing in k when $\phi(x) = x^2$. A similar impact of the activation function can be observed when considering the standard datasets MNIST and CIFAR-10 (third and fourth plot respectively). Additional experiments with different activation functions can be found in Appendix F.4.

5.4.1 Outline of the Argument

The proof of Theorem 5.5 can be divided into three steps. It will be convenient to define the shorthand

$$A(z) := \nabla_z \Phi_{\mathsf{RF}}(z)^\top \Phi_{\mathsf{RF}}^\top K_{\mathsf{RF}}^{-1} \in \mathbb{R}^{d\times n}. \tag{5.17}$$

Step 1. Fitting the noise lower bounds the sensitivity. First, we lower bound the sensitivity of our trained model with a quantity which does not depend on the labels and grows proportionally with the noise.

Theorem 5.6. *Let Assumptions 5.1, 5.2, 5.3 and 5.4 hold. Define $A(z)$ as in* (5.17)*. Then, we have*

$$\mathcal{S}_{\mathsf{RF}}(z) \geq \frac{\varepsilon}{2}\|z\|_2 \left\|A(z)\right\|_F, \tag{5.18}$$

with probability at least $1 - \exp\left(-c\left\|A(z)\right\|_F^2 / \left\|A(z)\right\|_{\mathsf{op}}^2\right)$ over ϵ, where c is a numerical constant.

The proof exploits the independence between $g(X)$ and ϵ and the Hanson-Wright inequality. The details are contained in Appendix F.2.1.

Theorem 5.6 removes the labels from the expression, and reduces the problem of estimating the robustness to characterizing the Frobenius norm of $A(z)$ (as long as $\left\|A(z)\right\|_F \gg \left\|A(z)\right\|_{\mathsf{op}}$, so that (5.18) holds with high probability). This is in line with Bubeck and Sellke [2021], Dohmatob [2022], which provide upper bounds on the robustness of models fitting the data below noise level.

Step 2. Splitting between interaction and kernel components. Next, we split the matrix $A(z)$ into two separate objects. The first is the *interaction matrix* given by

$$\mathcal{I}_{\mathsf{RF}}(z) := \nabla_z \Phi_{\mathsf{RF}}(z)^\top \tilde{\Phi}_{\mathsf{RF}}^\top, \tag{5.19}$$

where we use the shorthand $\tilde{\Phi}_{\mathsf{RF}} := \Phi_{\mathsf{RF}} - \mathbb{E}_X[\Phi_{\mathsf{RF}}] \in \mathbb{R}^{n\times k}$. The second is the *centered kernel* $\tilde{K}_{\mathsf{RF}} := \tilde{\Phi}_{\mathsf{RF}}\tilde{\Phi}_{\mathsf{RF}}^\top$, whose spectrum can be studied separately.

Theorem 5.7. *Let Assumptions 5.1, 5.2, 5.3 and 5.4 hold. Define $A(z)$ as in* (5.17)*. Then, we have*

$$\begin{aligned}
\left\|A(z)\right\|_F &\geq \lambda_{\max}^{-1}\left(\tilde{K}_{\mathsf{RF}}\right)\left\|\mathcal{I}_{\mathsf{RF}}(z)\right\|_F - C\sqrt{n+d}/d, \\
\left\|A(z)\right\|_F &\leq \lambda_{\min}^{-1}(\tilde{K}_{\mathsf{RF}})\left\|\mathcal{I}_{\mathsf{RF}}(z)\right\|_F + C\sqrt{n+d}/d,
\end{aligned} \tag{5.20}$$

which implies

$$\begin{aligned}
\left\|A(z)\right\|_F &\geq C_1\frac{d}{kn}\left\|\mathcal{I}_{\mathsf{RF}}(z)\right\|_F - C\sqrt{n+d}/d, \\
\left\|A(z)\right\|_F &\leq C_2 k^{-1}\left\|\mathcal{I}_{\mathsf{RF}}(z)\right\|_F + C\sqrt{n+d}/d,
\end{aligned} \tag{5.21}$$

with probability at least $1-\exp(-c\log^2 n)$ over X and V, where c, C, C_1 and C_2 are absolute constants.

Proof sketch. To prove the claim, first we characterize the extremal eigenvalues of the kernel $K_{\mathsf{RF}} := \Phi_{\mathsf{RF}}\Phi_{\mathsf{RF}}^\top$ and of its centered counterpart $\tilde{K}_{\mathsf{RF}} := \tilde{\Phi}_{\mathsf{RF}}\tilde{\Phi}_{\mathsf{RF}}^\top$, with $\tilde{\Phi}_{\mathsf{RF}} = \Phi_{\mathsf{RF}} - \mathbb{E}_X\Phi_{\mathsf{RF}}$,

see Appendix F.2.2. Then, we perform a delicate centering step on the matrix $A(z)$ defined in (5.17). Informally, we show that

$$\|\nabla_z \Phi_{\mathsf{RF}}(z)^\top \Phi_{\mathsf{RF}}^\top K_{\mathsf{RF}}^{-1}\|_F \approx \|\nabla_z \Phi_{\mathsf{RF}}(z)^\top \tilde{\Phi}_{\mathsf{RF}}^\top \tilde{K}_{\mathsf{RF}}^{-1}\|_F, \tag{5.22}$$

see Lemma F.19 for a precise statement. This is the key technical ingredient, since the rank-one components coming from the average feature matrix have a large operator norm, which would trivialize the lower bound on the sensitivity. More specifically, the centering leads to two rank-one terms which are successively removed via the Sherman-Morrison formula and a number of ad-hoc estimates, see Appendix F.2.3. Finally, the proof of (5.20)-(5.21) follows by combining (5.22) with the bounds on the smallest/largest eigenvalues of $\tilde{K}_{\mathsf{RF}}$, and it appears at the end of Appendix F.2.3. □

Theorem 5.7 unveils the crucial role played by the interaction matrix $\mathcal{I}_{\mathsf{RF}}(z)$ on the robustness of the model. This term is the only one containing the test sample z, and it depends on how the term $\nabla_z \Phi_{\mathsf{RF}}(z)$ *aligns* (or "interacts") with the centered feature map of the training data $\tilde{\Phi}_{\mathsf{RF}}$.

Step 3. Estimating $\|\mathcal{I}_{\mathsf{RF}}(z)\|_F$. Finally, we provide a precise estimate on the norm of the interaction matrix $\|\mathcal{I}_{\mathsf{RF}}(z)\|_F$. To do so, we assume that the test point z is independently sampled from the data distribution P_X.

Theorem 5.8. *Let Assumptions 5.1, 5.2, 5.3 and 5.4 hold. Let $z \sim P_X$ be sampled independently from the training set (X, Y), and $\mathcal{I}_{\mathsf{RF}}(z)$ be defined as in* (5.19)*. Then,*

$$\left| \|\mathcal{I}_{\mathsf{RF}}(z)\|_F - \mathbb{E}_\rho^2[\phi'(\rho)] \frac{k\sqrt{n}}{\sqrt{d}} \right| = o\left(\frac{k\sqrt{n}}{\sqrt{d}}\right), \tag{5.23}$$

with probability at least $1 - \exp(-c \log^2 k)$ over X, z and V, where c is an absolute constant.

Proof sketch. A direct calculation gives that

$$\|\mathcal{I}_{\mathsf{RF}}(z)\|_F^2 = \sum_{i=1}^n \left\| V^\top \mathsf{diag}\left(\phi'(Vz)\right) \tilde{\phi}(Vx_i) \right\|_2^2,$$

and we bound separately each term of the sum. Note that the three terms $V^\top$, $\mathsf{diag}\left(\phi'(Vz)\right)$ and $\tilde{\phi}(Vx_i)$ are correlated by the presence of V. However, V contributes to the last two terms only via a single projection (along the direction of z and x_i, respectively). Hence, the key idea is to use a Taylor expansion to split $\mathsf{diag}\left(\phi'(Vz)\right)$ and $\tilde{\phi}(Vx_i)$ into a component correlated with V and an independent one. The correlated components are computed exactly and the remaining independent term is shown to have a negligible effect. The details are in Appendix F.2.4. □

At this point, the proof of Theorem 5.5 follows by combining the results of Theorems 5.6-5.8. The details are in Appendix F.2.5.

We note that the results of Theorems 5.6-5.8 do not require the assumption $\mathbb{E}_{\rho\sim\mathcal{N}(0,1)}[\phi'(\rho)] \neq 0$. In fact, the role of this additional condition is clarified by the statement of Theorem 5.8: if $\mathbb{E}_{\rho\sim\mathcal{N}(0,1)}[\phi'(\rho)] \neq 0$, then $\|\mathcal{I}_{\mathsf{RF}}(z)\|_F$ is of order $\Theta(k\sqrt{n/d})$; otherwise, $\|\mathcal{I}_{\mathsf{RF}}(z)\|_F$ is drastically smaller. This provides an explanation to the qualitatively different behavior of the sensitivity for $\phi(x) = \tanh(x)$ and $\phi(x) = x^2$ displayed in Figure 5.1. To conclude, the interaction matrix appears to be the key quantity capturing the impact of the activation function on the robustness of the ERM solution.

5.5 Main Result for NTK Regression

In this section, we provide our law of robustness for the *NTK* model. In particular, we will show that, for a class of *even* activations, the sensitivity of NTK can be $O(1)$ and, therefore, the *model is robust* for some scaling of the parameters.

We consider the following two-layer neural network

$$f_{\mathsf{nn}}(x,w)=\sum_{i=1}^{k}\phi\left(W_{i:}^{1}x\right)-\sum_{i=1}^{k}\phi\left(W_{i:}^{2}x\right). \tag{5.24}$$

Here, the hidden layer contains $2k$ neurons; ϕ is an activation function applied component-wise; $W^1, W^2 \in \mathbb{R}^{k\times d}$ denote the first and second half of the weights of the hidden layer, respectively; for $j\in\{1,2\}$, $W_{i:}^{j}$ denotes the i-th row of W^j; the first k weights of the second layer are set to 1, and the last k weights to -1. We indicate with w the vector containing the parameters of this model, i.e., $w=[\mathsf{vec}(W^1),\mathsf{vec}(W^2)]\in\mathbb{R}^p$, with $p=2kd$. For convenience, we initialize the network so that its output is 0. This has been shown to be necessary to have a robust model after lazy training [Dohmatob and Bietti, 2022, Wang et al., 2022]. Specifically, we let $w_0=[\mathsf{vec}(W_0^1),\mathsf{vec}(W_0^2)]$ be the vector of the parameters at initialization, where we take $[W_0^1]_{i,j}\sim_{\text{i.i.d.}}\mathcal{N}(0,1/d)$ and $W_0^2=W_0^1$. Here, with a slight abuse of notation, we use the subscript 0 to refer to the initialization, and not to indicate matrix rows. This readily implies that $f_{\mathsf{nn}}(x,w_0)=0$, for all x.

Now, the *NTK regression model* takes the form

$$f_{\mathsf{NTK}}(x,\theta)=\Phi_{\mathsf{NTK}}(x)^{\top}\theta,\ \ \Phi_{\mathsf{NTK}}(x)=\nabla_w f_{\mathsf{nn}}(x,w)|_{w=w_0}. \tag{5.25}$$

Here, the vector trainable parameters is $\theta\in\mathbb{R}^p$, again with $p=2kd$, which is initialized with $\theta_0=w_0$. This is the same model considered by [Dohmatob and Bietti, 2022, Montanari and Zhong, 2020]. We remark that $f_{\mathsf{NTK}}(x,\theta)$ is equivalent to the linearization of $f_{\mathsf{nn}}(x,w)$ around the initial point w_0, see e.g. [Jacot et al., 2018, Bartlett et al., 2021].

An application of the chain rule gives

$$\Phi_{\mathsf{NTK}}(x)=\left[x\otimes\phi'(W_0^1x),-x\otimes\phi'(W_0^2x)\right]=:\left[\Phi'_{\mathsf{NTK}}(x),-\Phi'_{\mathsf{NTK}}(x)\right], \tag{5.26}$$

where the last equality follows from the fact that $W_0^1=W_0^2$, and the definition $\Phi'_{\mathsf{NTK}}(x):=x\otimes\phi'(W_0^1x)$. Thus, since $\theta_0=w_0=[\mathsf{vec}(W_0^1),\mathsf{vec}(W_0^2)]=[\mathsf{vec}(W_0^1),\mathsf{vec}(W_0^1)]$, we have

$$f_{\mathsf{NTK}}(x,\theta_0)=\Phi_{\mathsf{NTK}}(x)^{\top}\theta_0=\left[\Phi'_{\mathsf{NTK}}(x),-\Phi'_{\mathsf{NTK}}(x)\right]^{\top}\left[\mathsf{vec}(W_0^1),\mathsf{vec}(W_0^1)\right]=0. \tag{5.27}$$

This means that our model's output is 0 at initialization. Then, we can use (5.8) to express the sensitivity of the trained NTK regression model as

$$\mathcal{S}_{\mathsf{NTK}}(z)=\|z\|_2\left\|\nabla_z\Phi_{\mathsf{NTK}}(z)^{\top}\Phi_{\mathsf{NTK}}^{\top}K_{\mathsf{NTK}}^{-1}Y\right\|_2, \tag{5.28}$$

where $\Phi_{\mathsf{NTK}}\in\mathbb{R}^{n\times p}$ contains in its i-th row the feature map of the i-th training sample $\Phi_{\mathsf{NTK}}(x_i)$ and $K_{\mathsf{NTK}}=\Phi_{\mathsf{NTK}}\Phi_{\mathsf{NTK}}^{\top}$. The invertibility of the kernel will again follow from the proof of our main result, Theorem 5.12. Throughout this section, we make the following assumptions.

Assumption 5.9 (Activation function)**.** *The activation function ϕ satisfies the following properties:*

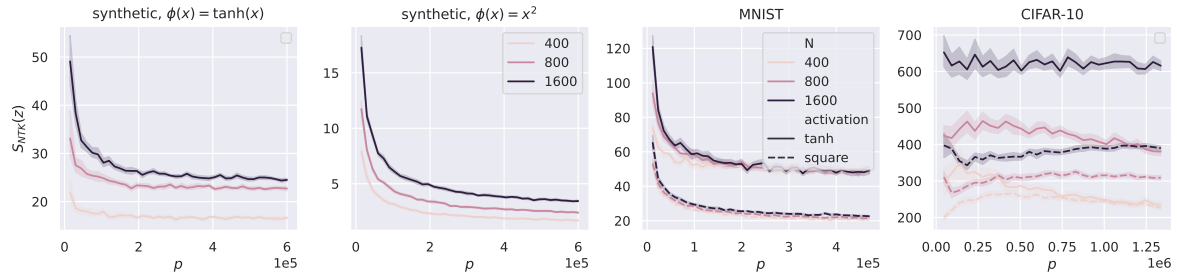


Figure 5.2: Sensitivity as a function of the number of parameters p, for an NTK model with synthetic data (first and second plot), and with inputs taken from two classes of the MNIST and CIFAR-10 datasets (third and fourth plot respectively). The rest of the setup is similar to that of Figure 5.1.

1. *ϕ is a non-linear, even function.*
2. *Its first order derivative ϕ' is an L-Lipschitz function.*

We restrict our analysis to an even activation function, as this requirement significantly simplifies the derivations. The impact of the activation on the robustness will also be discussed at the end of this section.

Assumption 5.10 (Minimum over-parameterization)**.**

$$n \log^8 n = o(kd). \tag{5.29}$$

Eq. (5.29) provides the weakest possible requirement on the number of parameters of the model capable to guarantee interpolation for generic data points, as it leads to a lower bound on the smallest eigenvalue of K_{NTK} [Bombari et al., 2022b].

Assumption 5.11 (High-dimensional data)**.**

$$k = O(d). \tag{5.30}$$

This requirement is purely technical, and we leave as a future work the interesting problem of characterizing the sensitivity of the NTK regression model when $d = o(k)$.

We will still use Assumption 5.1 on the data distribution P_X, and the only assumption on the labels is that $\|Y\|_2 = \Theta(\sqrt{n})$, which corresponds to the natural requirement that the output is $\Theta(1)$. At this point, we are ready to state our main result for NTK regression.

Theorem 5.12. *Let Assumptions 5.1, 5.9, 5.10 and 5.11 hold. Then, we have*

$$\mathcal{S}_{\mathsf{NTK}}(z) = O\left(\log k \left(1 + \sqrt{\frac{n}{k}}\right) \sqrt{\frac{n}{k}}\right), \tag{5.31}$$

with probability at least $1 - ne^{-c\log^2 k} - e^{-c\log^2 n}$ over X and w_0. Furthermore, when $n = O(k)$, (5.31) *simplifies to*

$$\mathcal{S}_{\mathsf{NTK}}(z) = O\left(\log k \sqrt{\frac{nd}{p}}\right), \tag{5.32}$$

where $p = 2dk$ denotes the number of parameters of the model.

Proof sketch. As for the RF model, we crucially separate the contribution of the interaction matrix and of the kernel of the model. Specifically, we upper bound the RHS of (5.28) with

$$\|z\|_2 \left\|\nabla_z \Phi_{\mathsf{NTK}}(z)^\top \Phi_{\mathsf{NTK}}^\top\right\|_{\mathsf{op}} \left\|K_{\mathsf{NTK}}^{-1}\right\|_{\mathsf{op}} \|Y\|_2 . \tag{5.33}$$

Now, each term in (5.33) is treated separately. First, by assumption, $\|Y\|_2 = \Theta(\sqrt{n})$. Next, we have that $\left\|K_{\mathsf{NTK}}^{-1}\right\|_{\mathsf{op}} = \lambda_{\min}^{-1}(K_{\mathsf{NTK}}) = O\left((dk)^{-1}\right)$, which can be deduced from [Bombari et al., 2022b], see Lemma F.28. Finally, the bound on $\left\|\nabla_z \Phi_{\mathsf{NTK}}(z)^\top \Phi_{\mathsf{NTK}}^\top\right\|_{\mathsf{op}}$ is computed explicitly through an application of the chain-rule (see Lemma F.30), and it critically depends on the data being 0-mean and on the activation being even. □

In words, Theorem 5.12 shows that, for even activations and k scaling super-linearly in n, $\mathcal{S}_{\mathsf{NTK}}(z) = O(1)$, namely, the NTK model is robust against adversarial perturbations. A few remarks are now in order.

Saturating the lower bound from the universal law of robustness. We highlight that our upper bound on the sensitivity in (5.32) exhibits the same scaling as the lower bound by [Bubeck and Sellke, 2021], thus saturating the universal law of robustness. A lower bound on the sensitivity of the same order is also provided by [Dohmatob, 2022], albeit restricted to the setting in which $k = \Theta(d)$ and $k = O(n)$. Note that Theorem 5.12 upper bounds the sensitivity of the model obtained by running gradient descent (over θ) on $f_{\mathsf{NTK}}(x, \theta)$. It is well known, see e.g. [Chizat et al., 2019], that a similar solution is obtained by running gradient descent (over w) on $f_{\mathsf{nn}}(x, w)$. This result holds after suitably rescaling the initialization of $f_{\mathsf{nn}}(x, w)$, and the rescaling does not affect our Theorem 5.12. As a result, Theorem 5.12 proves that a class of two-layer neural networks has sensitivity of order $\sqrt{nd/p}$ (up to logarithmic factors), thus resolving in the affirmative Conjecture 2 of [Bubeck et al., 2021].

Impact of the activation function and numerical results. Theorem 5.12 applies to even activation functions. At the technical level, the symmetry in ϕ directly "centers" the kernel in the expression (5.28) of the sensitivity, which largely simplifies our argument. We conjecture that this symmetry may in fact be fundamental to guarantee the robustness of the model. In fact, the first plot of Figure 5.2 shows that, for $\phi(x) = x^2$, the sensitivity keeps decreasing, as the number of neurons k (and, therefore, the number of parameters p) grows. In contrast, for $\phi(x) = \tanh(x)$, the sensitivity plateaus at a value which is an order of magnitude larger than for $\phi(x) = x^2$ (see the second plot of Figure 5.2). This difference is still visible even for real-world datasets (MNIST, CIFAR-10), as displayed in the third and fourth plot of Figure 5.2.

5.6 Conclusions

This chapter provides a precise and quantitative characterization of how the robustness of the solution obtained via empirical risk minimization depends on the model at hand: for random features and non-even activations, over-parameterization does not help, and we provide bounds tighter than the universal law of robustness proposed by [Bubeck and Sellke, 2021]; for NTK regression and even activations, the model is robust as soon as $p > dn$, i.e., the universal law of robustness is saturated. Numerical results on synthetic and standard datasets confirm the impact of the model on the robustness of the solution. The present contribution focuses on empirical risk minimization, which represents the cornerstone of deep learning training algorithms. In contrast, a number of recent papers has focused on adversarial training,

considering the trade-off between generalization and robust error, see e.g. [Raghunathan et al., 2019, Zhang et al., 2019, Dobriban et al., 2020, Carmon et al., 2019, Min et al., 2021, Hassani and Javanmard, 2024] and references therein. We conclude by mentioning that the technical tools developed in this chapter could be applied also to the models deriving from adversarial training, and we leave such an analysis as future work.

CHAPTER 6

Spurious Correlations in High Dimensional Regression: The Roles of Regularization, Simplicity Bias and Over-Parameterization

Learning models have been shown to rely on spurious correlations between non-predictive features and the associated labels in the training data, with negative implications on robustness, bias and fairness. In this chapter, we provide a statistical characterization of this phenomenon for high-dimensional regression, when the data contains a predictive *core* feature x and a *spurious* feature y. Specifically, we quantify the amount of spurious correlations $\mathcal{C}$ learned via linear regression, in terms of the data covariance and the strength λ of the ridge regularization. As a consequence, we first capture the simplicity of y through the spectrum of its covariance, and its correlation with x through the Schur complement of the full data covariance. Next, we prove a trade-off between $\mathcal{C}$ and the in-distribution test loss $\mathcal{L}$, by showing that the value of λ that minimizes $\mathcal{L}$ lies in an interval where $\mathcal{C}$ is increasing. Finally, we investigate the effects of over-parameterization via the random features model, by showing its equivalence to regularized linear regression. Our theoretical results are supported by numerical experiments on Gaussian, Color-MNIST, and CIFAR-10 datasets.

This chapter is based on the work published at ICML 2025: *Spurious correlations in high dimensional regression: The roles of regularization, simplicity bias and over-parameterization*.

6.1 Introduction

Machine learning systems have been shown to learn from patterns that are statistically correlated with the intended task, despite not being causally predictive [Geirhos et al., 2020, Xiao et al., 2021]. As a concrete example, a blue background in a picture might be positively correlated with the presence of a boat or a plane in the foreground, and while not being a predictive feature per se, a trained deep learning model could use this information to bias its prediction. In the literature, this statistical (but non causal) connection is referred to as a *spurious correlation* between a feature and the learning task. A recent and extensive line of research has investigated the extent to which deep learning models manifest this behavior

[Geirhos et al., 2019, Xiao et al., 2021] and has proposed different mitigation approaches [Sagawa et al., 2020a, Liu et al., 2021], given its implications to robustness, bias, and fairness [Zliobaite, 2015, Zhou et al., 2021a]. The phenomenon, also referred to as shortcut learning, is often attributed to the relative "simplicity" of spurious features [Geirhos et al., 2020, Shah et al., 2020, Hermann and Lampinen, 2020] and to the implicit bias of over-parameterized models toward learning simpler patterns [Belkin et al., 2019a, Rahaman et al., 2019, Kalimeris et al., 2019]. Consequently, the *core features* that are informative about the task (e.g., the boat in the foreground) may be neglected, as *spurious features* (e.g., the blue background) provide an easier shortcut to minimize the loss function.

Prior work has attempted to formalize the *simplicity bias* relying on boolean functions [Qiu et al., 2023], model-specific biases [Morwani et al., 2023], one-dimensional features [Shah et al., 2020] and their pairwise interactions [Pezeshki et al., 2021]. However, when considering high-dimensional natural data (e.g., the boat and its background in Figure 6.1), it remains unclear, based on these notions, what exactly makes the features easy or difficult to learn, and to what extent a trained model relies on spurious correlations. Furthermore, while Sagawa et al. [2020b] show that over-parameterization can exacerbate spurious correlations when re-weighting the objective on minority groups (e.g., boats with a green background), its effect on models trained via empirical risk minimization (ERM) is less understood. This is a critical point when additional group membership annotations are too expensive to obtain, and ERM is a key part of training [Liu et al., 2021, Ahmed et al., 2021].

This chapter tackles these issues: we provide a rigorous characterization of the statistical mechanisms behind learning spurious correlations in high-dimensional data, focusing on the solution obtained via ERM. Formally, we model the input sample z as composed by two distinct features, *i.e.*, $z = [x, y]$, where $x \in \mathbb{R}^d$ is the core feature and $y \in \mathbb{R}^d$ the spurious one. The first panel of Figure 6.1 provides an illustration with a boat in the foreground (x) and its blue background (y). Then, we quantify spurious correlations via the covariance $\mathcal{C}$ (see (6.3)) between the label g ("boat") and the model output given $\tilde{z} = [\tilde{x}, y]$ as input. Here, $\tilde{x}$ (a truck in the foreground) is a new core feature independent of everything else, see the second panel of Figure 6.1. Now, if $\mathcal{C}$ is positive, it means that the model is biased towards g only because of y, since $\tilde{x}$ is independent from x and g.

Figure 6.1: *Left two panels:* pictorial representation of the core (spurious) feature x (y) and an independent core feature $\tilde{x}$, taken from an image of a boat and a truck in the CIFAR-10 dataset. *Right two panels:* examples from a binary Color-MNIST dataset, where the labels correspond to the number shapes, and the zeros (ones) are colored in blue (red) with probability $(1 + \alpha)/2$.

More precisely, we provide a sharp, non-asymptotic characterization of $\mathcal{C}$ for linear regression (Theorem 6.3). Armed with such a characterization, we then:

• Interpret $\mathcal{C}$ via upper bounds on its magnitude (Proposition 6.4). This highlights the role of the regularization strength and of the data covariance via *(i)* its Schur complement with respect to the covariance of the core feature x, and *(ii)* the covariance of the spurious feature y. Specifically, we link the smallest eigenvalue of the Schur complement to the strength of the correlation between y and x, and the largest eigenvalue of the spurious covariance to the simplicity of y.

• Prove a trade-off between $\mathcal{C}$ and the test loss (Proposition 6.6), which implies that spurious

correlations can be beneficial to performance when learning in-distribution. Specifically, we show that the optimal regularization minimizing the test loss lies in an interval where $\mathcal{C}$ is positive and monotonically increasing.

• Investigate the role of over-parameterization via a *random features* (RF) model. Specifically, we show that the RF model is equivalent to linear regression with an effective regularization that depends on the over-parameterization (Theorem 6.10). This allows to leverage the earlier analysis on regularized linear regression to quantify spurious correlations in over-parameterized, non-linear models.

Throughout this chapter, the theoretical results are supported by numerical experiments on synthetic Gaussian data, Color-MNIST, and CIFAR-10, which validates our analysis even in settings not strictly following the modeling choices. Our code is publicly available at the GitHub repository `simone-bombari/spurious-correlations`.

6.2 Related work

Spurious correlations. Learning from spurious correlations in a training dataset is rather common [Geirhos et al., 2019, Arjovsky et al., 2020, Geirhos et al., 2020, Sagawa et al., 2020a, Xiao et al., 2021, Singla and Feizi, 2022] and it has unwanted consequences, e.g., lack of robustness towards domain shift, prediction bias and compromised algorithmic fairness [Zliobaite, 2015, Geirhos et al., 2019, Zhou et al., 2021a, Veitch et al., 2021, Seo et al., 2022]. Thus, multiple mitigation approaches have been proposed, with [Sagawa et al., 2020a, Zhang et al., 2021] or without [Liu et al., 2021, Ahmed et al., 2021] available annotations. Specifically, Tiwari and Shenoy [2023] exploit the difference in the features learned at different layers of a deep neural network; Izmailov et al. [2022], Kirichenko et al. [2023] re-train the last layer of the ERM solution to adapt the features to the distribution shift; and Chang et al. [2021a], Plumb et al. [2022] mitigate the problem via data augmentation.

Simplicity bias. Recent work has shown that deep learning models have a bias towards learning from "easier" patterns [Belkin et al., 2019a, Rahaman et al., 2019, Kalimeris et al., 2019]. In shortcut learning, this property is formalized in different ways across the literature. The difficulty of a feature is defined in terms of the minimum complexity of a network that learns it by Hermann and Lampinen [2020] and in terms of the smallest amount of linear segments that separate different classes by Shah et al. [2020]. Moayeri et al. [2022] connect the simplicity to the position and size of the features in an image. Morwani et al. [2023] define the simplicity bias in 1-hidden layer neural networks via the rank of a projection operator that does not alter them substantially, and they focus on a dataset generated via an independent features model learned via the NTK. The NTK is also used to analyze gradient starvation [Pezeshki et al., 2021] and feature availability [Hermann et al., 2024], regarded as explanations of the simplicity bias. Qiu et al. [2023] focus on parity functions and staircases, analyzing the learning dynamics of features having different complexity.

High-dimensional regression. The test loss of linear regression when the input dimension d scales proportionally with the sample size n has been characterized precisely both in-distribution [Hastie et al., 2022, Cheng and Montanari, 2024] and under covariate shift [Yang et al., 2023, Mallinar et al., 2024, Song et al., 2024]. Furthermore, Montanari et al. [2025], Chang et al. [2021b], Han and Xu [2023] have studied the distribution of the ERM solution via the convex Gaussian min-max Theorem [Thrampoulidis et al., 2015]. Specifically, our work builds on the non-asymptotic characterization provided by Han and Xu [2023].

In contrast with linear regression where the number of parameters equals the input dimension, random features models [Rahimi and Recht, 2007] capture the effects of over-parameterization, as the number of parameters is independent of d and n. Some related literature on this model is discussed in Chapter 3, This model has been also considered in the setting where the data has two dependent components [Loureiro et al., 2021], although it was used to study the training and generalization error when only partial information is available. Furthermore, its robustness to distribution shift is studied also by Tripuraneni et al. [2021], Lee et al. [2023]. The equivalence between an over-parameterized RF model and a regularized linear one has also been studied in detail [Goldt et al., 2022, 2020, Hu and Lu, 2022, Montanari and Saeed, 2022]. However, existing rigorous results show the equivalence at the level of training and test error. In contrast, we are interested in the covariance defined in (6.3) and, for this reason, we prove an equivalence at the level of the predictor (Theorem 6.10).

6.3 Preliminaries

Notation. Given a vector v, we denote by $\|v\|_2$ its Euclidean norm. Given a matrix A, we denote by $\mathrm{tr}(A)$ and $\|A\|_{\mathrm{op}}$ its trace and operator (spectral) norm. Given a symmetric matrix A, we denote by $\lambda_{\min}(A)$ ($\lambda_{\max}(A)$) its smallest (largest) eigenvalue. We denote by λ (without a sub-script) the ℓ_2 regularization term in ridge regression. All complexity notations $\Omega(\cdot)$, $O(\cdot)$, $\omega(\cdot)$, $o(\cdot)$ and $\Theta(\cdot)$ are understood for large data size n, input dimension d, and number of parameters p. We indicate with $C, c > 0$ numerical constants, independent of n, d, p, whose value may change from line to line.

Setting. We consider supervised learning with n training samples $\{(z_1, g_1), \dots, (z_n, g_n)\}$ and labels defined by a (not necessarily deterministic) function of the inputs $g_i = f^*(z_i)$, where $z_i \in \mathbb{R}^{2d}$ denotes the i-th training input and $g_i \in \mathbb{R}$ the corresponding label. Input samples are composed by two distinct parts (or *features*), *i.e.*, $z_i^\top = [x_i^\top, y_i^\top]$, with $x_i, y_i \in \mathbb{R}^d$, and they are sampled i.i.d. from the distribution $\mathcal{P}_{XY}$. We further denote with $\mathcal{P}_X$ ($\mathcal{P}_Y$) the marginal distribution of the x_i-s (y_i-s). The features x and y have the same dimension d to ease the presentation. We opted to focus on the setting $z = [x, y]$ due to its connection with the motivating example of images with background, but the analysis could be extended to other cases, such as $z = x + y$, briefly discussed in Appendix G.3.

We focus on the setting where the labels g_i depend only on x_i, *i.e.*, $g_i = f^*(z_i) = f_x^*(x_i)$ for some (not necessarily deterministic) function f_x^*. Hence, y_i is independent from g_i, after conditioning on x_i. We highlight that the independence between y_i and g_i is conditional on x_i, and that the covariance between y_i and x_i is in general non-zero. We refer to y_i as the *spurious feature* of the i-th sample, and to x_i as its *core feature*. As an example, x_i may represent the main object in an image and y_i the (not necessarily independent) background, see Figure 6.1.

In this setup, the training data is used to learn $f^*(z)$ through a parametric model $f(\theta, z)$ via regularized empirical risk minimization (ERM). Specifically, we perform the following optimization in parameter space:

$$\hat{\theta} = \arg\min_\theta \left(\frac{1}{n} \sum_{i=1}^{n} \ell\left(f(\theta, z_i), g_i\right) + \lambda \|\theta\|_2^2 \right), \tag{6.1}$$

for some regularization term $\lambda \geq 0$, where ℓ is a loss function[1]. We define the test loss associated to the model $f(\hat{\theta}, \cdot)$ as

$$\mathcal{L}(\hat{\theta}) = \mathbb{E}_{z \sim \mathcal{P}_{XY},\, g = f^*(z)} \left[\ell \left(f(\hat{\theta}, z), g \right) \right]. \tag{6.2}$$

We note that this test loss is in distribution.

Spurious correlations. We express the extent to which a model $f(\hat{\theta}, \cdot)$ learns spurious correlations between the spurious feature y and the label g as

$$\mathcal{C}(\hat{\theta}) = \mathrm{Cov} \left(f \left(\hat{\theta}, [\tilde{x}^\top, y^\top]^\top \right), g \right), \tag{6.3}$$

where the covariance is computed on the probability space of $[x^\top, y^\top]^\top \sim \mathcal{P}_{XY}$, $g = f_x^*(x)$ and of the independent core feature $\tilde{x} \sim \mathcal{P}_X$. In words, $\mathcal{C}(\hat{\theta})$ expresses how the output of the model $f(\hat{\theta}, \cdot)$ evaluated on an out-of-distribution sample $[\tilde{x}^\top, y^\top]^\top$ (where the two features are sampled independently from the marginal distributions $\mathcal{P}_X$ and $\mathcal{P}_Y$) correlates to the label associated to the in-distribution sample $g = f^*(z) = f_x^*(x)$. We highlight that, if the model $f(\hat{\theta}, \cdot)$ *does not rely* on the spurious feature y, then $\mathcal{C}(\hat{\theta}) = 0$ as x and $\tilde{x}$ are independent.

We note that (6.3) formalizes to the regression setting the definition in the survey [Ye et al., 2024] and the fairness metric in [Zliobaite, 2015] (when interpreting y as the protected variable). Furthermore, in the context of classification, it can connected to the worst group accuracy [Sagawa et al., 2020a,b]. In fact, our definition (6.3) is also related to the *out-of-distribution* test loss: assuming $\mathbb{E}[f(\hat{\theta}, [\tilde{x}^\top, y])^2] = \mathbb{E}[f_x^*(\tilde{x})^2] = 1$ and $\mathbb{E}[f(\hat{\theta}, [\tilde{x}^\top, y])] = \mathbb{E}[f_x^*(\tilde{x})] = 0$ for simplicity and considering a quadratic loss, we have

$$\mathbb{E}_{\tilde{x}, y} \left[\left(f(\hat{\theta}, [\tilde{x}^\top, y]) - f_x^*(\tilde{x}) \right)^2 \right] \geq 2 - 2\sqrt{1 - \mathcal{C}(\hat{\theta})^2}. \tag{6.4}$$

This implies that an increase in $\mathcal{C}(\hat{\theta})$ hurts the performance of the model when core and spurious features are sampled independently (and, thus, the model is tested out-of-distribution). The proof of (6.4) is in Appendix G.4.

6.4 Precise analysis for linear regression

To study $\mathcal{C}(\cdot)$ as defined in (6.3), we focus on a high-dimensional *linear regression* model, *i.e.*,

$$f_{\mathsf{LR}}(\theta, z) = z^\top \theta, \tag{6.5}$$

where $\theta \in \mathbb{R}^{2d}$. The data also follows a linear model, *i.e.*,

$$g_i = z_i^\top \theta^* + \epsilon_i = x_i^\top \theta_x^* + \epsilon_i, \tag{6.6}$$

where $\theta^* \in \mathbb{R}^{2d}$, $\theta_x^* \in \mathbb{R}^d$, and ϵ_i is label noise. The second equality in (6.6) implies that $\theta^* = [\theta_x^{*\top}, \mathbf{0}_d^\top]^\top$, where $\theta_x^*, \mathbf{0}_d \in \mathbb{R}^d$ and each entry of $\mathbf{0}_d$ is 0. We set $\|\theta^*\|_2 = \|\theta_x^*\|_2 = 1$ and let the ϵ_i-s be i.i.d. (and independent from the z_i-s), mean-0, sub-Gaussian, with variance $\sigma^2 > 0$.

[1] In general, existence and uniqueness of $\hat{\theta}$ depend on the choice of the model $f(\theta, z)$, the loss function ℓ and the regularization term λ. For the purposes of this chapter, we will precisely define $\hat{\theta}$ for linear regression (Section 6.4) and for random features (Section 6.6).

We introduce the shorthands $Z = [z_1^\top, \dots, z_n^\top]^\top \in \mathbb{R}^{n \times 2d}$, $G = [g_1, \dots, g_n]^\top \in \mathbb{R}^n$, and $\mathcal{E} = [\epsilon_1, \dots, \epsilon_n]^\top \in \mathbb{R}^n$ to indicate the data matrix, the labels, and the noise vector respectively. Then, using a quadratic loss, (6.1) reads

$$\hat{\theta}_{\mathsf{LR}}(\lambda) = \arg\min_\theta \left(\frac{1}{n} \|Z\theta - G\|_2^2 + \lambda \|\theta\|_2^2 \right), \tag{6.7}$$

which admits the unique solution

$$\hat{\theta}_{\mathsf{LR}}(\lambda) = \left(Z^\top Z + n\lambda I \right)^{-1} Z^\top G, \tag{6.8}$$

for $\lambda > 0$ and, if $Z^\top Z$ is invertible, also for $\lambda = 0$.

Assumption 6.1 (Data distribution)**.** *$\{z_i\}_{i=1}^n$ are n i.i.d. samples from the multivariate, mean-0, Gaussian distribution $\mathcal{P}_{XY}$, such that its covariance $\Sigma := \mathbb{E}\left[zz^\top\right] \in \mathbb{R}^{2d \times 2d}$ is invertible, with $\lambda_{\max}(\Sigma) = O(1)$, $\lambda_{\min}(\Sigma) = \Omega(1)$, and $\operatorname{tr}(\Sigma) = 2d$.*

This requirement could be relaxed to having sub-Gaussian data. We focus on the Gaussian case for simplicity, deferring the discussion on the generalization to Appendix G.1.2.

Warm-up: no regularization ($\lambda = 0$). Our first result concerns the amount of spurious correlations learned in the un-regularized setting.

Proposition 6.2. *Let $\lambda = 0$ and $Z^\top Z \in \mathbb{R}^{2d \times 2d}$ be invertible*[2]*. Let $\mathcal{C}(\hat{\theta}_{\mathsf{LR}}(0))$ be the amount of spurious correlations learned by the model $f_{\mathsf{LR}}(\hat{\theta}_{\mathsf{LR}}(0), \cdot)$. Then, we have that*

$$\mathbb{E}_{\mathcal{E}}[\mathcal{C}(\hat{\theta}_{\mathsf{LR}}(0))] = 0. \tag{6.9}$$

Furthermore, if Assumption 6.1 holds and $n = \omega(d)$,

$$|\mathcal{C}(\hat{\theta}_{\mathsf{LR}}(0))| = \mathcal{O}(\log d / \sqrt{d}), \tag{6.10}$$

with probability at least $1 - 2\exp(-c \log^2 d)$ over Z and $\mathcal{E}$, where c is an absolute constant.

In words, $f_{\mathsf{LR}}(\hat{\theta}_{\mathsf{LR}}(0), \cdot)$ *does not learn* any spurious correlation between the spurious feature y and the label g. This is also clear from Figure 6.2, where we report in red the value of $\mathcal{C}(\hat{\theta}_{\mathsf{LR}}(\lambda))$, which approaches 0 as λ becomes small.

The idea of the argument is to write explicitly the solution

$$\hat{\theta}_{\mathsf{LR}}(0) = \left(Z^\top Z \right)^{-1} Z^\top G = \theta^* + \left(Z^\top Z \right)^{-1} Z^\top \mathcal{E}, \tag{6.11}$$

where in the second step we separate the ground truth θ^* (which does not capture any dependence on y) from a term only depending on the label noise, which is mean-0 and independent from y. This directly gives (6.9). Then, the bound in (6.10) is obtained via standard concentration results on $\lambda_{\min}\left(Z^\top Z \right)$. The details are in Appendix G.1.

[2]Under Assumption 6.1, this holds with probability 1 for $n \geq 2d$.

General case with regularization ($\lambda > 0$). Setting a regularizer $\lambda > 0$ often reduces the test loss, see the black curve in Figure 6.2. However, it also leads to non-trivial spurious correlations, and our main result provides a non-asymptotic characterization of this phenomenon.

Theorem 6.3. *Let Assumption 6.1 hold, $n = \Theta(d)$ and $\mathcal{C}(\hat{\theta}_{\mathsf{LR}}(\lambda))$ be the amount of spurious correlations learned by the model $f_{\mathsf{LR}}(\hat{\theta}_{\mathsf{LR}}(\lambda), \cdot)$ for $\lambda > 0$. Denote by $P_y \in \mathbb{R}^{2d \times 2d}$ the projector on the last d elements of the canonical basis in $\mathbb{R}^{2d}$, and set*

$$\mathcal{C}^{\Sigma}(\lambda) := \theta^{*\top} \Sigma \left(\Sigma + \tau(\lambda) I\right)^{-1} P_y \Sigma \theta^*, \tag{6.12}$$

where $\tau := \tau(\lambda)$ is implicitly defined as the unique positive solution of

$$1 - \frac{\lambda}{\tau} = \frac{1}{n} \operatorname{tr}\left(\left(\Sigma + \tau I\right)^{-1} \Sigma\right). \tag{6.13}$$

Then, for every $t \in (0, 1/2)$,

$$\mathbb{P}_{Z,\mathcal{E}}\left(\left|\mathcal{C}(\hat{\theta}_{\mathsf{LR}}(\lambda)) - \mathcal{C}^{\Sigma}(\lambda)\right| \geq t\right) \leq C d \exp\left(-d t^4 / C\right),$$

where C is an absolute constant.

In words, Theorem 6.3 guarantees that $|\mathcal{C}(\hat{\theta}_{\mathsf{LR}}(\lambda)) - \mathcal{C}^{\Sigma}(\lambda)| = o(1)$ with high probability (e.g., setting $t = d^{-1/5}$). Thus, for large d, n, we can theoretically analyze $\mathcal{C}(\hat{\theta}_{\mathsf{LR}}(\lambda))$ via the deterministic quantity $\mathcal{C}^{\Sigma}(\lambda)$, which, as highlighted by (6.12), depends on θ^*, the covariance of the data Σ, and the regularization λ via the parameter $\tau(\lambda)$ introduced in (6.13). We further analyze the object $\mathcal{C}^{\Sigma}(\lambda)$ in Section 6.5 that immediately follows.

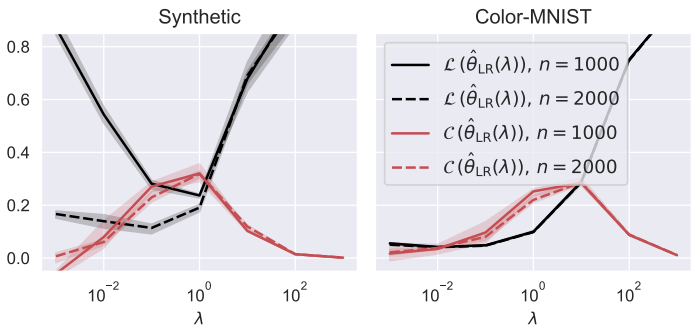


Figure 6.2: Test loss $\mathcal{L}(\hat{\theta}_{\mathsf{LR}}(\lambda))$ (black) and spurious correlations $\mathcal{C}(\hat{\theta}_{\mathsf{LR}}(\lambda)$ (red) as a function of the regularization term λ for two values of the number of samples n. *Left:* synthetic Gaussian dataset, with $d = 400$ (additional details in Appendix G.5); *right:* binary Color-MNIST dataset with correlation $\sqrt{1-\alpha^2} = 0.25$ between color and digit (see Figure 6.1).

Note that, since $\hat{\theta}_{\mathsf{LR}}(\lambda)$ is given by (6.8), when $\lambda > 0$ it cannot be decomposed as $\theta^* + \left(Z^\top Z\right)^{-1} Z^\top \mathcal{E}$ (as in the proof of Proposition 6.2 for $\lambda = 0$). Thus, we rely on the non-asymptotic characterization of $\hat{\theta}_{\mathsf{LR}}(\lambda)$ recently provided by [Han and Xu, 2023]. In particular, their analysis focuses on the proportional regime $n = \Theta(d)$ (as opposed to the regime $n = \omega(d)$, considered in (6.10)), and it allows to provide concentration bounds on a certain family of low-dimensional functions of $\hat{\theta}_{\mathsf{LR}}(\lambda)$, which includes $\mathcal{C}$ as defined in (6.3). The details are in Appendix G.1.

6.5 Roles of regularization and simplicity bias

We now interpret $\mathcal{C}^{\Sigma}(\lambda)$, which characterizes the spurious correlations via Theorem 6.3, in terms of the data covariance Σ and the regularization λ. To do so, we express Σ as

$$\Sigma = \left(\begin{array}{c|c} \Sigma_{xx} & \Sigma_{xy} \\ \hline \Sigma_{yx} & \Sigma_{yy} \end{array}\right), \tag{6.14}$$

where the block $\Sigma_{xx} = \mathbb{E}_{x\sim\mathcal{P}_X}\left[xx^\top\right] \in \mathbb{R}^{d\times d}$ ($\Sigma_{yy} = \mathbb{E}_{y\sim\mathcal{P}_Y}\left[yy^\top\right] \in \mathbb{R}^{d\times d}$) denotes the covariance of the core (spurious) feature sampled from its marginal distribution. The off-diagonal blocks are $\Sigma_{xy} = \Sigma_{yx}^\top = \mathbb{E}_{[x^\top,y^\top]^\top\sim\mathcal{P}_{XY}}\left[xy^\top\right] \in \mathbb{R}^{d\times d}$. Let us denote by

$$S_x^{\Sigma} := \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy} \tag{6.15}$$

the Schur complement of Σ with respect to the top-left $d \times d$ block Σ_{xx}. In our setting, S_x^{Σ} offers a helpful statistical interpretation. In fact, for multivariate Gaussian data, it corresponds to the conditional covariance of y given x, *i.e.*,

$$S_x^{\Sigma} = \mathrm{Cov}\left(y|x=\bar{x}\right) = \mathbb{E}_{y|x=\bar{x}}\left[\left(y-\mathbb{E}_{y|x=\bar{x}}[y]\right)\left(y-\mathbb{E}_{y|x=\bar{x}}[y]\right)^\top\right]. \tag{6.16}$$

Therefore, the spectrum of S_x^{Σ} describes the degree of dependence between y and x: on the one hand, if its eigenvalues are small, the feature y is close to be determined by the knowledge of the feature x (*i.e.*, y is highly correlated with x); on the other hand, if its eigenvalues are large, the two features tend to be independent.

As an intuitive example, let us interpret classification for the binary Color-MNIST dataset (see Figure 6.1) as a 2-dimensional problem: the core feature is $x = +1$ (-1) if the digit on the image is 1 (0); and the spurious feature is $y = +1$ (-1) if the color is red (blue). If the digits and the colors have a correlation of α, then $\Sigma_{xx} = \Sigma_{yy} = 1$ and $\Sigma_{xy} = \Sigma_{yx} = \alpha$. Thus, $S_x^{\Sigma} = 1 - \alpha^2$, which becomes 0 if $|\alpha| = 1$ (full correlation), and it is maximized in the setting of independence between colors and parity ($\alpha = 0$).

At this point, leveraging the decomposition of Σ in (6.14) and the Schur complement in (6.15), we provide the following bounds on $\mathcal{C}^{\Sigma}(\lambda)$, which are proved in Appendix G.1.

Proposition 6.4. *Let $\mathcal{C}^{\Sigma}(\lambda)$ and S_x^{Σ} be defined in* (6.12) *and* (6.15)*, respectively. Then,*

$$\left|\mathcal{C}^{\Sigma}(\lambda)\right| \leq \min\left(\left\|\Sigma_{yx}\right\|_{\mathsf{op}}, \frac{\lambda_{\max}\left(\Sigma\right)^2}{\tau(\lambda)}, \tau(\lambda)\sqrt{\mathrm{Var}(g)-\sigma^2}\frac{\lambda_{\max}\left(\Sigma_{yy}\right)-\lambda_{\min}\left(S_x^{\Sigma}\right)}{\lambda_{\min}\left(S_x^{\Sigma}\right)\sqrt{\lambda_{\min}\left(\Sigma_{xx}\right)}}\right). \tag{6.17}$$

We discuss the three upper bounds in (6.17) below.

(i): $|\mathcal{C}^{\Sigma}(\lambda)| \leq \|\Sigma_{yx}\|_{\mathsf{op}}$. The off-diagonal blocks $\Sigma_{yx} = \mathbb{E}[yx^\top]$ and $\Sigma_{xy} = \Sigma_{yx}^\top$ describe the correlation between y and x. In the limit case $\|\Sigma_{yx}\|_{\mathsf{op}} = 0$, we have that x and y are uncorrelated and, therefore, $\mathcal{C}^{\Sigma}(\lambda) = 0$, as there is no spurious correlation that the model can learn.

(ii): $|\mathcal{C}^{\Sigma}(\lambda)| \leq \lambda_{\max}\left(\Sigma\right)^2/\tau(\lambda)$. From (6.13), one obtains that $\tau(\lambda) \to \infty$ as $\lambda \to \infty$. Thus, the bound implies that $\mathcal{C}^{\Sigma}(\lambda)$ approaches 0 as λ grows large. This captures the intuition that, when the regularization λ is large, the minimization in (6.7) is biased towards solutions with small norm and, therefore, the output of the model is small, which drives to 0 the spurious

correlations as defined in (6.3). The behavior is confirmed by Figure 6.2: $|\mathcal{C}(\hat{\theta}_{\mathsf{LR}}(\lambda))|$ is decreasing for large values of λ and it eventually vanishes; at the same time, large values of λ make the output of the model small, which in turn increases the test loss $\mathcal{L}(\hat{\theta}_{\mathsf{LR}}(\lambda))$.

(iii): The third bound in (6.17) can be rewritten as

$$\frac{|\mathcal{C}^{\Sigma}(\lambda)|\sqrt{\lambda_{\min}\left(\Sigma_{xx}\right)}}{\sqrt{\mathrm{Var}(g)-\sigma^2}} \leq \tau(\lambda)\frac{\lambda_{\max}\left(\Sigma_{yy}\right)-\lambda_{\min}\left(S_x^{\Sigma}\right)}{\lambda_{\min}\left(S_x^{\Sigma}\right)}, \tag{6.18}$$

where we have isolated on the LHS the terms depending on the covariance of the core feature x ($\sqrt{\lambda_{\min}\left(\Sigma_{xx}\right)}$) and on the scaling of the labels ($\sqrt{\mathrm{Var}(g)-\sigma^2}$). We now discuss the dependence of the RHS of (6.18) w.r.t. *(a)* $\tau(\lambda)$, *(b)* $\lambda_{\min}\left(S_x^{\Sigma}\right)$, and *(c)* $\lambda_{\max}\left(\Sigma_{yy}\right)$. As for *(a)*, we note that $\mathcal{C}^{\Sigma}(\lambda)$ approaches 0 for small values of λ. In fact, the RHS of (6.13) is smaller or equal to $2d/n$; thus, if we consider $2d < n$, we also get $\tau \leq \lambda\left(1-2d/n\right)^{-1}$, which implies $\tau(\lambda) \to 0$ as $\lambda \to 0$. This is in agreement with Proposition 6.2, which handles the case without regularization, and also with the numerical experiments of Figure 6.2. As for *(b)*, we note that the bound is decreasing with $\lambda_{\min}\left(S_x^{\Sigma}\right)$. This is in agreement with the earlier discussion on how the spectrum of the Schur complement S_x^{Σ} measures the degree of independence between the spurious feature y and the core feature x. Finally, as for *(c)*, we note that the bound is increasing with $\lambda_{\max}\left(\Sigma_{yy}\right)$, which is connected below to the *simplicity* of the spurious feature y. The increasing (decreasing) trend of $\mathcal{C}^{\Sigma}(\lambda)$ w.r.t. $\lambda_{\max}\left(\Sigma_{yy}\right)$ ($\lambda_{\min}\left(S_x^{\Sigma}\right)$) is clearly displayed in Figure 6.3 for Gaussian data.

The connection between $\lambda_{\max}\left(\Sigma_{yy}\right)$ and the *simplicity bias* of ERM can be illustrated via our initial image recognition example. The (spurious) background feature is intuitively an easy pattern to learn from the model: the pixels corresponding to the spurious feature behave consistently across the training data. This in turn skews the spectrum of Σ_{yy}, which has few dominant directions with eigenvalues much larger than the others. Note that this interpretation is similar to the model-dependent definition of simplicity in [Morwani et al., 2023]. An empirical verification is provided in Figure 6.4, where we consider the CIFAR-10 dataset, restricted to the "boat" and "truck" classes. Before training a regression model, we whiten up to some level the background feature (as defined in Figure 6.1) to make it harder to learn, see the right side of Figure 6.4. Then, for different levels of whitening, we report $\mathcal{C}(\hat{\theta}_{\mathsf{LR}}(\lambda))$ as a function of $\lambda_{\max}\left(\Sigma_{yy}\right)$. We normalize $\lambda_{\max}\left(\Sigma_{yy}\right)$ by the trace $\mathrm{tr}(\Sigma_{yy})$ to exclude the size of the pattern from our experiment[3]. The red curve shows an increasing trend analogous to that displayed in Figure 6.3 for Gaussian data: small values of $\lambda_{\max}\left(\Sigma_{yy}\right)$

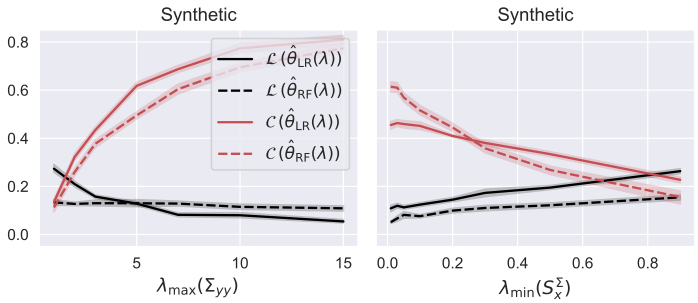


Figure 6.3: Test loss $\mathcal{L}(\hat{\theta}_{\mathsf{LR/RF}}(\lambda))$ (black) and spurious correlations $\mathcal{C}(\hat{\theta}_{\mathsf{LR/RF}}(\lambda))$ (red) as a function of $\lambda_{\max}\left(\Sigma_{yy}\right)$ (left) and $\lambda_{\min}\left(S_x^{\Sigma}\right)$ (right) on a synthetic Gaussian dataset, for both linear regression and random features, with $\lambda = 1$ (additional details in Appendix G.5).

[3] If y has 0-mean, then $\mathbb{E}\|y\|_2^2 = \mathrm{tr}(\Sigma_{yy})$, i.e., the trace captures the size of the pattern.

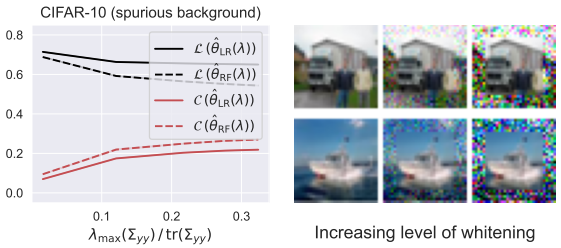


Figure 6.4: Test loss $\mathcal{L}(\hat{\theta}_{\mathsf{LR/RF}}(\lambda))$ (black) and spurious correlations $\mathcal{C}(\hat{\theta}_{\mathsf{LR/RF}}(\lambda)$ (red) as a function of $\lambda_{\max}(\Sigma_{yy})\,/\,\mathrm{tr}(\Sigma_{yy})$ on a CIFAR-10 dataset for different levels of whitening (details on the whitening process in Appendix G.5). We restrict to the classes "boat" and "truck" ($n = 10000$) and consider both linear regression and random features, with $\lambda = 1$.

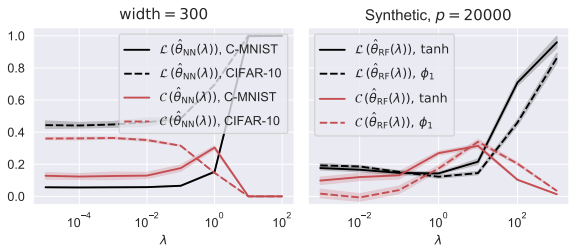


Figure 6.5: Test loss $\mathcal{L}(\hat{\theta}_{\mathsf{NN/RF}}(\lambda))$ (black) and spurious correlations $\mathcal{C}(\hat{\theta}_{\mathsf{NN/RF}}(\lambda)$ (red) as a function of λ. *Left:* 2-layer fully connected ReLU network, trained on the binary color(C)-MNIST and CIFAR-10 (boats and trucks). *Right:* RF model with $\tanh$ and $\phi_1 = h_1 + 0.1\,h_3$ activation. Implementation details are in Appendix G.5.

correspond to significant whitening and, hence, to small spurious correlations, as predicted by Proposition 6.4.

We remark that our results concern the parameter $\hat{\theta}$ as defined in (6.1), which can be interpreted as the convergence point of an optimization algorithm such as gradient descent. This perspective differs from the prior work of [Pezeshki et al., 2021, Qiu et al., 2023], which focus on how spurious correlations evolve during training. On the other hand, in linear regression, solving the gradient flow equation $\mathrm{d}\theta = -\nabla_\theta \mathcal{L}_\lambda(\theta)\mathrm{d}t$, where $\mathcal{L}_\lambda(\theta)$ is defined as the argument of the $\arg\min$ in (6.1), gives

$$\theta(t) = \left(1 - \exp\left(-2\left(X^\top X/n + \lambda I\right)t\right)\right)\hat{\theta}. \tag{6.19}$$

Thus, the components of $\hat{\theta}$ aligned with the top eigen-spaces of $X^\top X$ converge earlier than the others. Hence, from a dynamical point of view, if $X^\top X \sim n\Sigma$, our results suggest that spurious features are learned faster the easier they are.

Trade-off between $\mathcal{L}(\hat{\theta}_{\mathsf{LR}}(\lambda))$ and $\mathcal{C}(\hat{\theta}_{\mathsf{LR}}(\lambda))$. Figure 6.2 shows that there is an interval of values for the regularization ($\lambda \sim 10^{-1}$) where the test loss $\mathcal{L}(\hat{\theta}_{\mathsf{LR}}(\lambda))$ is decreasing in λ, while the spurious correlations $\mathcal{C}(\hat{\theta}_{\mathsf{LR}}(\lambda))$ are increasing. This evidence suggests a natural trade-off between these two quantities, mediated by λ. To theoretically capture such trade-off, we first provide a non-asymptotic concentration bound for $\mathcal{L}(\hat{\theta}_{\mathsf{LR}}(\lambda))$.

Proposition 6.5. *Let Assumption 6.1 hold, $n = \Theta(d)$ and $\mathcal{L}(\hat{\theta}_{\mathsf{LR}}(\lambda))$ be the in-distribution test loss of the model $f_{\mathsf{LR}}(\hat{\theta}_{\mathsf{LR}}(\lambda), \cdot)$ for $\lambda > 0$. Set*

$$\mathcal{L}^{\Sigma}(\lambda) := \frac{\sigma^2 + \tau(\lambda)^2 \left\|(\Sigma + \tau(\lambda)I)^{-1}\,\Sigma^{1/2}\theta^*\right\|_2^2}{1 - \frac{\mathrm{tr}\left((\Sigma+\tau(\lambda)I)^{-2}\Sigma^2\right)}{n}}, \tag{6.20}$$

where $\tau(\lambda)$ is defined via (6.13)*. Then, for every $t \in (0, 1/2)$,*

$$\mathbb{P}_{Z,\mathcal{E}}\left(\left|\mathcal{L}(\hat{\theta}_{\mathsf{LR}}(\lambda)) - \mathcal{L}^{\Sigma}(\lambda)\right| \geq t\right) \leq Cd\exp\left(-dt^4/C\right),$$

where C is an absolute constant.

In words, Proposition 6.5 guarantees that $|\mathcal{L}(\hat{\theta}_{\mathsf{LR}}(\lambda)) - \mathcal{L}^{\Sigma}(\lambda)| = o(1)$ with high probability. Its proof is an adaptation of Theorem 3.1 in [Han and Xu, 2023], and the details are in Appendix G.1.

Armed with the non-asymptotic bounds of Theorem 6.3 and Proposition 6.5, we characterize the trade-off between $\mathcal{L}(\hat{\theta}_{\mathsf{LR}}(\lambda))$ and $\mathcal{C}(\hat{\theta}_{\mathsf{LR}}(\lambda))$ by studying the monotonicity of $\mathcal{L}^{\Sigma}(\lambda)$ and $\mathcal{C}^{\Sigma}(\lambda)$.

Proposition 6.6. *Let $\mathcal{C}^{\Sigma}(\lambda)$ and $\mathcal{L}^{\Sigma}(\lambda)$ be defined as in* (6.12) *and* (6.20)*. Then, if $2d < n$, we have that $\mathcal{L}^{\Sigma}(\lambda)$ is monotonically decreasing in a right neighborhood of $\lambda = 0$, and there exists $\lambda_{\mathcal{L}} > 0$ such that $\mathcal{L}^{\Sigma}(\lambda)$ is monotonically increasing for $\lambda \geq \lambda_{\mathcal{L}}$. Furthermore, if $\Sigma_{xx} = I$, then $\mathcal{C}^{\Sigma}(\lambda)$ is non-negative and there exists $\lambda_{\mathcal{C}}$ such that $\mathcal{C}^{\Sigma}(\lambda)$ is monotonically increasing for $\lambda \leq \lambda_{\mathcal{C}}$. Finally, as long as*

$$\frac{2d}{n} \leq \frac{\lambda_{\min}(\Sigma)}{4} \min\left(1, \frac{2\lambda_{\max}(\Sigma)/\sigma^2}{\left(\lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma) + 1\right)^2}\right), \tag{6.21}$$

we have that $\lambda_{\mathcal{C}} \geq \lambda_{\mathcal{L}}$.

In words, Proposition 6.6 shows that $\mathcal{C}^{\Sigma}(\lambda)$ grows with λ at least until the regularization equals a value $\lambda_{\mathcal{C}}$. For example, in Figure 6.2, $\lambda_{\mathcal{C}} \sim 1$ for a Gaussian data and $\lambda_{\mathcal{C}} \sim 10$ for Color-MNIST. Furthermore, in this interval, $\mathcal{L}^{\Sigma}(\lambda)$ is initially decreasing and then increasing as $\lambda \geq \lambda_{\mathcal{L}}$. These trends in turn imply that the optimal value $\lambda^*_{\mathcal{L}}$ that minimizes the test loss is s.t. $\lambda^*_{\mathcal{L}} \in (0, \lambda_{\mathcal{C}}]$ – an interval where the spurious correlations are strictly positive and increasing.

The proof of Proposition 6.6 (whose details are in Appendix G.1) relies on the monotonicity of $\tau(\lambda)$ in λ, and the last statement follows from showing that $\tau(\lambda_{\mathcal{C}}) \geq \lambda_{\min}\left(S_x^{\Sigma}\right) \geq \lambda_{\min}(\Sigma) \geq \tau(\lambda_{\mathcal{L}})$. The upper bound on $2d/n$ in (6.21) is required to prove that $\lambda_{\min}(\Sigma) \geq \tau(\lambda_{\mathcal{L}})$ and, due to Assumption 6.1, it is implied by taking $n = \omega(d)$. We note that the latter scaling holds in standard datasets, e.g., MNIST ($n = 6 \cdot 10^4$, $2d \approx 2 \cdot 10^3$ when considering the 3 color channels) and CIFAR-10 ($n = 5 \cdot 10^4$, $2d \approx 3 \cdot 10^3$).

We conclude the section by noting that learning spurious correlations can be beneficial to minimize the (in-distribution) test loss. In fact, the spurious features in y are effectively correlated with the labels, due to their correlation with the core feature x, and hence they can be helpful at prediction time. This phenomenon is numerically supported by Figures 6.3 and 6.4, where for a fixed value of λ, easier spurious features (or higher correlations) generate both higher values of $\mathcal{C}(\hat{\theta}_{\mathsf{LR}}(\lambda))$ and lower values of $\mathcal{L}(\hat{\theta}_{\mathsf{LR}}(\lambda))$. In words, while a blue background cannot strictly predict the label "boat", it is a useful feature in prediction as long as the boats in the test data tend to have a blue background.

The same conclusion does not hold for the out-of-distribution test loss, where the features x and y are sampled independently: the lower bound in (6.4) increases with $\mathcal{C}$. Figure G.1 in Appendix G.4 provides an additional numerical validation of this statement.

6.6 Role of Over-parameterization

Our analysis has so far focused on linear regression, highlighting the role of data covariance and regularization. However, moving to complex predictive models, such as neural networks,

may lead to differences in the degree to which spurious correlations are learned. As an example, in the left panel of Figure 6.5, we train an over-parameterized two-layer neural network on the binary Color-MNIST and CIFAR-10 datasets, for different values of the regularizer λ. While for high values of λ the results are qualitatively similar to the ones in Figure 6.2, a striking difference is that spurious correlations remain significant even when there is little to no regularization (i.e., $\lambda \approx 0$), in sharp contrast with Proposition 6.2. We also note that the phenomenon is in line with previous empirical work [Sagawa et al., 2020a].

We bridge the gap between linear regression and over-parameterized models by focusing on *random features*:

$$f_{\mathsf{RF}}(\theta, z) = \phi(Vz)^\top \theta, \tag{6.22}$$

where V is a $p \times 2d$ matrix s.t. $V_{i,j} \sim_{\text{i.i.d.}} \mathcal{N}(0, 1/(2d))$, and ϕ is an activation applied component-wise. The number of parameters of this model is p, as V is a fixed random matrix and $\theta \in \mathbb{R}^p$ contains trainable parameters. The scaling of input data ($\operatorname{tr}(\Sigma) = 2d$) and the variance of the entries of V guarantee that the pre-activations of the model (*i.e.*, the entries of the vector $Vz \in \mathbb{R}^p$) are of constant order.

We consider the ERM in (6.1) with a quadratic loss

$$\hat{\theta}_{\mathsf{RF}}(\lambda) = \arg\min_\theta \left(\frac{1}{n} \|\Phi\theta - G\|_2^2 + \lambda \|\theta\|_2^2 \right), \tag{6.23}$$

where we set $\Phi := [\phi(Vz_1), \dots, \phi(Vz_n)]^\top \in \mathbb{R}^{n \times p}$. When $\lambda = 0$, if $\Phi\Phi^\top$ is invertible, the minimization above does not necessarily have a unique solution. In that case, we set $\hat{\theta}_{\mathsf{RF}}(0)$ to be the solution obtained via gradient descent with 0 initialization, which corresponds to the min-norm interpolator (see equation (33) in [Bartlett et al., 2021]). Then[4], we can write, for $\lambda \geq 0$,

$$\hat{\theta}_{\mathsf{RF}}(\lambda) = \Phi^\top \left(\Phi\Phi^\top + n\lambda I \right)^{-1} G. \tag{6.24}$$

Assumption 6.7 (Activation function)**.** *The activation $\phi : \mathbb{R} \to \mathbb{R}$ is a non-linear, odd, Lipschitz function, such that its first Hermite coefficient $\mu_1 \neq 0$.*

This choice is motivated by theoretical convenience and is similar to the one considered in [Hu and Lu, 2022]. We believe that our result can be extended to a more general setting, as the ones in [Mei and Montanari, 2022, Mei et al., 2021], with a more involved analysis. We refer to [O'Donnell, 2014] for background on Hermite coefficients.

Assumption 6.8 (Over-parameterization)**.** *We let p grow s.t. $p = \omega\left(n \log^4 n\right)$ and $\log p = \Theta(\log n)$.*

This requires the width of the model (and, hence, its number of parameters) to grow faster (by at least a poly-log factor) than the number of training samples.

Finally, our requirements on the data are less restrictive than those coming from Assumption 6.1.

Assumption 6.9 (Data distribution, less restrictive)**.** *$\{z_i\}_{i=1}^n$ are n i.i.d. samples from a mean-0, Lipschitz concentrated distribution $\mathcal{P}_{XY}$, with covariance Σ s.t. $\operatorname{tr}(\Sigma) = 2d$. Furthermore, the labels g_i are i.i.d. sub-Gaussian random variables.*

[4] $\Phi\Phi^\top$ is proved to be invertible with high probability in Lemma G.10.

Note that the labels g_i are not required to follow a linear model $g_i = z_i^\top \theta^* + \epsilon_i$. The Lipschitz concentration property (see Appendix B for details) corresponds to data having well-behaved tails, it includes the distributions considered in Assumption 6.1, as well as the uniform distribution on the sphere or the hypercube [Vershynin, 2018], and it is a common requirement in the related literature [Nguyen et al., 2021b, Bubeck and Sellke, 2021, Bombari et al., 2022b].

Theorem 6.10. *Let Assumptions 6.7, 6.8, and 6.9 hold, $n = \Theta(d)$, and $z \in \mathbb{R}^{2d}$ be sampled from a distribution satisfying Assumption 6.9, not necessarily with the same covariance as $\mathcal{P}_{XY}$, independent from everything else. Let $f_{\mathsf{RF}}(\hat{\theta}_{\mathsf{RF}}(\lambda), z)$ be the RF model defined in* (6.22) *with $\hat{\theta}_{\mathsf{RF}}(\lambda)$ given by* (6.24)*, and $f_{\mathsf{LR}}(\hat{\theta}_{\mathsf{LR}}(\tilde{\lambda}), z)$ be the linear regression model defined in* (6.5) *with $\hat{\theta}_{\mathsf{LR}}(\tilde{\lambda})$ given by* (6.8)*. Then, for $\lambda \geq 0$,*

$$\begin{aligned} &\left| f_{\mathsf{RF}}(\hat{\theta}_{\mathsf{RF}}(\lambda), z) - f_{\mathsf{LR}}(\hat{\theta}_{\mathsf{LR}}(\tilde{\lambda}), z) \right| \\ &\qquad = O\left(\frac{d^{1/4} \log d}{p^{1/4}} + \frac{\log^{3/2} d}{d^{1/8}} \right) = o(1), \end{aligned} \tag{6.25}$$

with probability at least $1 - C\sqrt{d}\log^2 d/\sqrt{p} - C\log^3 d/d^{1/4}$, where the effective regularization *$\tilde{\lambda}$ is given by*

$$\tilde{\lambda} = \frac{2\tilde{\mu}^2 d}{\mu_1^2 n} + \frac{2d}{\mu_1^2 p}\lambda, \tag{6.26}$$

and $\tilde{\mu}^2 = \sum_{k\geq 2} \mu_k^2$, with μ_k denoting the k-th Hermite coefficient of ϕ.

In words, Theorem 6.10 shows that the over-parameterized RF model, when evaluated on a new test sample (not necessarily from the same distribution as the input data), is asymptotically equivalent to linear regression with regularization $\tilde{\lambda}$ given by (6.26). In particular, even in the ridgeless case ($\lambda = 0$), the RF model is equivalent to linear regression with strictly positive regularization. Thus, we expect the presence of spurious correlations, just like in Figure 6.5, since $\mathcal{C}(\hat{\theta}_{\mathsf{RF}}(0))$ approaches $\mathcal{C}^{\Sigma}(\tilde{\lambda})$ with $\tilde{\lambda} > 0$. Notably, the effective regularization $\tilde{\lambda}$ depends on the activation ϕ via its Hermite coefficients, and it increases with the ratio $\tilde{\mu}^2/\mu_1^2$.

We point out some differences between Theorem 6.10 and earlier work on the equivalence of random features with regularized linear regression [Goldt et al., 2020, 2022, Hu and Lu, 2022, Montanari and Saeed, 2022]. First, as mentioned in Section 6.2, existing results prove equivalence of training and test loss, which does not imply either the equivalence of the covariance in (6.3) nor the point-wise guarantee of (6.25). Second, existing results focus on the regime where p and n are proportional, while Assumption 6.8 requires $p = \omega\left(n \log^4 n\right)$. This reflects on the third difference which is in the proof technique: existing results use Lindeberg method, while our strategy (summarized below) relies on concentration tools.

Proof sketch: The proof builds on 3 core steps.

Step 1: We show that, with high probability,

$$\left\| \frac{\Phi\Phi^\top}{p} - \mu_1^2 \frac{ZZ^\top}{2d} - \tilde{\mu}^2 I \right\|_{\mathsf{op}} = o\left(\frac{\lambda_{\min}\left(\Phi\Phi^\top\right)}{p} \right),$$

which is a consequence of the concentration of $\Phi\Phi^\top$ to its expectation with respect to V, then expressed in terms of the Hermite coefficients of ϕ (Lemmas G.9 and G.10).

Step 2: In Lemma G.12 we upper bound the term $\|\mathbb{E}_z[\tilde{\phi}(Vz)\tilde{\phi}(Vz)^\top]\|_{\text{op}}$ (where we set $\tilde{\phi}(\cdot) := \phi(\cdot) - \mu_1(\cdot)$), which is then used to show that

$$\left|\phi(Vz)^\top \hat{\theta}_{\text{RF}} - \mu_1 z^\top V^\top \hat{\theta}_{\text{RF}}\right| = o(1),$$

with high probability, due to Markov inequality. This means that the non-linear component of $\phi(Vz)$ has a negligible effect on the output (see (G.123) in Lemma G.14).

Step 3: Using a similar intuition, we use matrix Bernstein inequality (Lemma G.13) to show that $\|(\Phi - \mu_1 Z V^\top)V\|_{\text{op}}$ is small with high probability, so that

$$\left|\mu_1 z^\top V^\top \left(\Phi^\top - \mu_1 V Z^\top\right)\left(\Phi\Phi^\top + n\lambda I\right)^{-1} G\right| = o(1),$$

i.e., the non-linear component of $\Phi^\top$ in (6.24) is also negligible (see (G.119)).

Finally, by combining *Step 2* and *Step 3* with standard concentration arguments, we conclude that

$$\left|\phi(Vz)^\top \hat{\theta}_{\text{RF}} - \mu_1^2 p \frac{z^\top Z^\top}{2d}(\Phi\Phi^\top + n\lambda I)^{-1} G\right| = o(1),$$

and the thesis follows from *Step 1* and Woodbury matrix identity for the inverse. □

The right panel of Figure 6.5 presents the test loss $\mathcal{L}(\hat{\theta}_{\text{RF}}(\lambda))$ (in black) and the spurious correlations $\mathcal{C}(\hat{\theta}_{\text{RF}}(\lambda))$ (in red) for two activation functions: $\tanh$ and $\phi_1 = h_1 + 0.1h_3$, where h_1 and h_3 denote the first and third Hermite polynomials, respectively. Notice that this gives $\tilde{\mu}^2/\mu_1^2 \sim 0.1$ for $\tanh$, $\tilde{\mu}^2/\mu_1^2 \sim 0.01$ for ϕ_1, and we take $d = 400$ and $n = 2000$. As expected, $\mathcal{C}(\hat{\theta}_{\text{RF}}(0)) > 0$ for the $\tanh$ activation function, since $\tilde{\lambda} \sim 0.05$ (which matches the corresponding value in Figure 6.2). On the other hand, $\mathcal{C}(\hat{\theta}_{\text{RF}}(0)) \sim 0$ for the activation ϕ_1, since $\tilde{\lambda} \sim 0.005$. As λ grows, $\mathcal{C}(\hat{\theta}_{\text{RF}}(\lambda))$ goes to 0 faster for the $\tanh$ activation function (which has higher $\tilde{\lambda}$), as predicted by the second upper bound in Proposition 6.4. These results match the qualitative behavior of the 2-layer neural network trained on the Color-MNIST dataset (left panel).

6.7 Conclusions

This chapter provides a rigorous study of spurious correlations in high-dimensional regression models, with the purpose of connecting the statistical foundation of this phenomenon with its practical intuition. Specifically, we translate the problem into characterizing the deterministic object $\mathcal{C}^\Sigma(\lambda)$, which allows us to quantitatively capture the roles of ridge regularization, data covariance and over-parameterization.

An interesting future direction is to employ our principled approach to design algorithms that go beyond ERM to mitigate spurious correlations. In that regard, let us mention the usage of multiple models with different ridge regularization or early stopping (connected to ridge penalties, see Raskutti et al. [2014]), in order to generate additional supervision. Here, having a quantitative control on which features (core or spurious) are learned by which model would allow to optimally use the extra labels (e.g., for up-weighting minority groups, as in Liu et al. [2021]).

CHAPTER 7

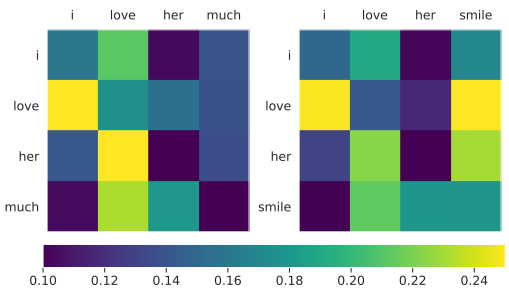
Towards Understanding the Word Sensitivity of Attention Layers: A Study via Random Features

Understanding the reasons behind the exceptional success of transformers requires a better analysis of why attention layers are suitable for NLP tasks. In particular, such tasks require predictive models to capture contextual meaning which often depends on one or few words, even if the sentence is long. This chapter studies this key property, dubbed *word sensitivity* (WS), in the prototypical setting of random features. We show that attention layers enjoy high WS, namely, there exists a vector in the space of embeddings that largely perturbs the random attention features map. The argument critically exploits the role of the $\mathrm{softmax}$ in the attention layer, highlighting its benefit compared to other activations (e.g., ReLU). In contrast, the WS of standard random features is of order $1/\sqrt{n}$, n being the number of words in the textual sample, and thus it decays with the length of the context. We then translate these results on the word sensitivity into generalization bounds: due to their low WS, random features provably cannot learn to distinguish between two sentences that differ only in a single word; in contrast, due to their high WS, random attention features have higher generalization capabilities. We validate our theoretical results with experimental evidence over the BERT-Base word embeddings of the imdb review dataset.

This chapter is based on the work published at ICML 2024: *Towards understanding the word sensitivity of attention layers: A study via random features*.

7.1 Introduction

Deep learning theory has provided a quantitative description of phenomena routinely occuring in state-of-the-art models, such as double-descent [Nakkiran et al., 2020, Mei and Montanari, 2022], benign overfitting [Bartlett et al., 2020, Belkin, 2021], and feature learning [Ba et al., 2022, Damian et al., 2022]. However, most existing works focus on architectures given by the composition of matrix multiplications and non-linearities, which model e.g. fully connected and convolutional layers. In contrast, the recent impressive results achieved by large language models [Brown et al., 2020, Bubeck et al., 2023] are largely attributed to the introduction of the transformer architecture [Vaswani et al., 2017], which are in turn based on attention layers



Prompt	Output
Reply with "Yes" if the review I will provide you is positive, and "No" otherwise. Review: Sorry, gave it a 1, which is the rating I give to movies on which I walk out or fall asleep.	No
Reply with "Yes" if the review I will provide you is negative, and "No" otherwise. Review: Sorry, gave it a 1, which is the rating I give to movies on which I walk out or fall asleep.	Yes

Figure 7.1: *Left.* Average attention scores for the word embeddings of the two sentences "I love her much" and "I love her smile". The embeddings are computed with the BERT-Base model, the scores are averaged over the 12 heads and displayed without the `[CLS]` token. *Right.* Output of the Llama2-7b-chat model for two prompts differing only in a single word.

[Bahdanau et al., 2014, Kim et al., 2017]. Hence, isolating the unique features of the attention mechanism stands out as a critical challenge to understand the success of transformers, thus paving the way to the principled design of large language models.

Recent work tackling this problem characterizes the sample complexities required by simplified attention models [Edelman et al., 2022, Fu et al., 2023]. However, learning is limited to a specific set of targets, such as sparse functions [Edelman et al., 2022] or functions of the correlations between the first query token and key tokens [Fu et al., 2023]. This chapter takes a different perspective and starts from the simple empirical observation that *one or few words can change the meaning of a sentence*. Think, for example, to the pair of sentences "I love her much" and "I love her smile", where replacing a single word alters the meaning of the text. This is captured by the BERT-Base model [Devlin et al., 2019], which shows a different attention score pattern over these two sentences (see Figure 7.1, left). Another example can be found in the table on the right of Figure 7.1, where changing one word in the prompt would require a well-behaved model (in this case, Llama2-7b [Touvron et al., 2023]) to modify its output. In general, language models need to have a high *word sensitivity* to capture semantic changes when just a single word is modified in the context, which motivates the following question:

Do attention layers have a larger word sensitivity than fully connected architectures?

To formalize the problem, we represent the textual data with $X \in \mathbb{R}^{N \times d}$, where the N rows represent the word embeddings in $\mathbb{R}^d$. Then, we define the word sensitivity (WS) in (7.4), as a measure of how changing a row in X modifies the embedding of a given feature map. We focus our study on the prototypical setting of *random features*, i.e., where the weights of the layers are random. In particular, we consider *(i)* the *random features* (RF) map [Rahimi and Recht, 2007], defined in (7.1), and *(ii)* the *random attention features* (RAF) map [Fu et al., 2023], defined in (7.3). The former models a fully connected architecture, while the latter captures the structure typical of attention layers.

Our contributions can be summarized as follows:

- Theorem 7.2 shows that an RF map has *low* WS, specifically of order $1/\sqrt{N}$, where N is the context length. This means that changing a single word has a negligible effect on the output of the map. In fact, to have a significant effect, one needs to change a

constant fraction of the words, see Remark 7.4. Furthermore, increasing the depth of the architecture does not help with the word sensitivity, as shown by Theorem 7.3.

- Our main result, Theorem 7.4, shows that a RAF map has *high* WS, specifically of constant order which does not depend on the length of the context. This means that changing even a single word can have a significant effect on the output of the map, regardless of the context length. The argument critically exploits the role of the $\mathrm{softmax}$ in the attention layer, and numerical simulations show its advantages compared to other activations (e.g., ReLU).
- Section 7.6 exploits the bounds on the word sensitivity to characterize the generalization error when the data changes meaning after modifying a single word in the context. In particular, we consider generalized linear models trained on RF and RAF embeddings, and we establish whether a fine-tuned or retrained model can learn to distinguish between two samples that differ only in one row. While the answer is provably negative for random features (Theorems 7.6 and 7.7), random attention features are capable of generalizing.

Most of our technical contributions require no distributional assumptions on the data, and the generality of our findings is confirmed by numerical results on the BERT-Base embeddings of the imdb dataset [Maas et al., 2011], and on the pre-trained BERT-Base model itself. Our code is publicly available at the GitHub repository `simone-bombari/attention-sensitivity`.

7.2 Related work

Fully connected layers. Several mathematical models have been proposed to understand phenomena occurring for fully connected architectures. Two prototypical examples are the RF and the NTK models, discussed in Chapters 2 and 3. Another popular approach involves a mean-field analysis of the neurons of a fully connected 2-layer neural network [Mei et al., 2018, Sirignano and Spiliopoulos, 2020, Chizat and Bach, 2018, Rotskoff and Vanden-Eijnden, 2018, Javanmard et al., 2020a, Shevchenko et al., 2021]. Deep random models have also been considered by Hanin [2019], Bosch et al. [2023], Schröder et al. [2023].

Attention layers. Attention layers [Bahdanau et al., 2014, Kim et al., 2017] and transformer architectures [Vaswani et al., 2017] have attracted significant interest from the theoretical community: Yun et al. [2020], Ben-Shaul et al. [2023] study their approximation capabilities; Edelman et al. [2022], Trauger and Tewari [2023] provide norm-based generalization bounds; Tian et al. [2023] analyze the training dynamics; Wu et al. [2023] provide optimization guarantees; Jelassi et al. [2022], Li et al. [2023] focus on computer vision tasks and Oymak et al. [2023] on prompt-tuning. The study of the attention mechanism is approached through the lens of associative memories by Bietti et al. [2023], Cabannes et al. [2024]. More closely related to our setting is the recent work by Fu et al. [2023], which compares the sample complexity of random attention features with that of random features. We highlight that Fu et al. [2023] focus on an attention layer with ReLU activation, while we unveil the critical role of the $\mathrm{softmax}$.

Sensitivity of neural networks. We informally use the term sensitivity to express how a perturbation of the input changes the output of the model. Previous work explored various

mathematical formulations of this concept (e.g., the input-output Lipschitz constant) in the context of both robustness [Weng et al., 2018, Bubeck and Sellke, 2021] and generalization [Bartlett et al., 2017]. Sensitivity is generally referred to as an *undesirable* property, motivating research on models that reduce it [Miyato et al., 2018, Prach and Lampert, 2022]. In this chapter, however, high sensitivity represents a *desirable* attribute, as it reflects the ability of the model to capture the role of individual words in a long context. This is a stronger requirement than having large Lipschitz constant, hence earlier results on the matter [Kim et al., 2021] cannot be applied.

7.3 Preliminaries

We consider a sequence of N tokens $\{x_i\}_{i=1}^N$, with $x_i \in \mathbb{R}^d$ for every i, where d denotes the token embedding dimension, and N the context length. These tokens altogether represent the textual sample $X = [x_1, \dots, x_N]^\top \in \mathbb{R}^{N\times d}$. We denote by $\mathrm{flat}(X) \in \mathbb{R}^D$, with $D = Nd$, the flattened (or vectorized) version of X. Given a vector x, $\|x\|_2$ is its Euclidean norm. Given a matrix A, $\|A\|_2 := \|\mathrm{flat}(A)\|_2 \equiv \|A\|_F$ is its Frobenius norm. We indicate with e_i the i-th element of the canonical basis, and denote $[N] := \{1, \dots, N\}$. All the complexity notations $\Omega(\cdot)$, $\omega(\cdot)$, $\mathcal{O}(\cdot)$, $o(\cdot)$ and $\Theta(\cdot)$ are understood for sufficiently large context length N, token embedding dimension d, number of neurons k, and number of input samples n. We indicate with $C, c > 0$ numerical constants, independent of N, d, k, n. Throughout the chapter, we make the following assumption, which is easily achieved by pre-processing the raw data.

Assumption 7.1 (Normalization of token embedding)**.** *For every token x_i, we assume $\|x_i\|_2 = \sqrt{d}$.*

Random Features (RF). A fully connected layer with random weights is commonly referred to as a *random features* map [Rahimi and Recht, 2007]. The map acts from a vector of covariates to a feature space $\mathbb{R}^k$, where k denotes the number of neurons. Thus, we flatten the context $\mathrm{flat}(X) \in \mathbb{R}^D$ before feeding it in the layer, and the RF map $\varphi_{\mathsf{RF}} : \mathbb{R}^{N\times d} \to \mathbb{R}^k$ takes the form

$$\varphi_{\mathsf{RF}}(X) = \phi(V\,\mathrm{flat}(X)), \tag{7.1}$$

where $\phi : \mathbb{R} \to \mathbb{R}$ is a non-linearity applied component-wise and $V \in \mathbb{R}^{k\times D}$ is the random features matrix, with $V_{i,j} \sim_{\text{i.i.d.}} \mathcal{N}(0, 1/D)$. This scaling of the variance of V ensures that the entries of $V\,\mathrm{flat}(X)$ have unit variance, as $\|\mathrm{flat}\,X\|_2 = \sqrt{D}$ by Assumption 7.1. We later consider a similar model with several random layers, referred to as *deep random features* (DRF) model and recently considered by [Bosch et al., 2023, Schröder et al., 2023].

Random Attention Features (RAF). We consider a single-head sequence-to-sequence self-attention layer without biases $\varphi_{\mathsf{QKV}}(X)$ [Vaswani et al., 2017], given by

$$\varphi_{\mathsf{QKV}}(X) = \mathrm{softmax}\left(\frac{XW_Q^\top W_K X^\top}{\sqrt{d'}}\right) XW_V^\top, \tag{7.2}$$

where the $\mathrm{softmax}$ is applied row-wise and defined as $\mathrm{softmax}(s)_i = e^{s_i}/\sum_j e^{s_j}$; $W_Q, W_K, W_V \in \mathbb{R}^{d'\times d}$ are respectively the queries, keys and values weight matrices. As in previous related work [Fu et al., 2023], we simplify the above expression with the re-parameterization $W := W_Q^\top W_K \in \mathbb{R}^{d\times d}$, and by removing the values weight matrix. This is for convenience of presentation, and our results can be generalized to the case where queries, keys and values are

random independent features, see Remark 7.5 at the end of Section 7.5. Thus, we define the *random attention features* layer $\varphi_{\mathsf{RAF}} : \mathbb{R}^{N\times d} \to \mathbb{R}^{N\times d}$ as

$$\varphi_{\mathsf{RAF}}(X) = \mathrm{softmax}\left(\frac{XWX^\top}{\sqrt{d}}\right) X, \tag{7.3}$$

where $W_{i,j} \sim_{\text{i.i.d.}} \mathcal{N}(0, 1/d)$. We refer to the argument of the $\mathrm{softmax}$ with the shorthand $S(X) := XWX^\top/\sqrt{d} \in \mathbb{R}^{N\times N}$. The scaling of the variance of W ensures that the entries of $S(X)$ have unit variance. Differently from [Fu et al., 2023], we do not consider the biased initialization discussed in [Trockman and Kolter, 2023], designed to make the diagonal elements of W positive in expectation. As remarked in [Trockman and Kolter, 2023], this initialization is aimed at replicating the final attention scores of vision transformers, and its utility on language models is less discussed. Furthermore, we remark that our simplified model does not include any of the architectural tweaks introduced to allow transformer models to process longer contexts [Bertsch et al., 2023, Beltagy et al., 2020], as it aims to capture the specific properties of a single attention layer.

Word sensitivity (WS). The word sensitivity (WS) measures how the embedding of a given mapping $\varphi(X)$ is impacted by a change in a single word. Given a sample $X = [x_1, \dots, x_N]^\top \in \mathbb{R}^{N\times d}$, the *perturbed* sample $X^i(\Delta)$ is obtained from X by setting its i-th row to $x_i + \Delta$ and keeping the remaining rows the same. Here, $\Delta \in \mathbb{R}^d$ denotes the perturbation of the i-th token, and its magnitude does not exceed the scaling of Assumption 7.1, i.e., $\|\Delta\|_2 \le \sqrt{d}$. Formally, given a mapping φ, we are interested in

$$\mathcal{S}_\varphi(X) = \sup_{i\in[N],\, \|\Delta\|_2\le\sqrt{d}} \frac{\|\varphi(X^i(\Delta)) - \varphi(X)\|_2}{\|\varphi(X)\|_2}. \tag{7.4}$$

In words, $\mathcal{S}_\varphi(X)$ denotes the highest relative change of $\varphi(X)$, upon changing a single token in the input. Our goal is to study how $\mathcal{S}_\varphi(X)$ behaves for the RF and RAF models, with respect to the context length N. Informally, if $\mathcal{S}_\varphi(X) = o(1)$ for any X, the mapping φ has *low word sensitivity*. On the contrary, if $\mathcal{S}_\varphi(X) = \Omega(1)$, φ has *high word sensitivity*.

We remark that, in language models, tokens are elements of a discrete vocabulary. However, working directly in the embedding space (as in definition (7.4)) is common in the theoretical literature [Kim et al., 2021], and critical steps of practical algorithms for language models also take place in the embedding space [Ebrahimi et al., 2018, Shin et al., 2020, Zou et al., 2023].

The definition of WS recalls similar notions of sensitivity in the context of adversarial robustness [Dohmatob and Bietti, 2022, Bubeck and Sellke, 2021, Wu et al., 2021, Bombari et al., 2023], as well as the literature that designs adversarial prompting schemes for language models [Ebrahimi et al., 2018, Guo et al., 2021, Zou et al., 2023]. However, in contrast with the definition in (7.4), these works consider a *trained* model and, therefore, the results implicitly depend on the training dataset. Our analysis of the WS dissects the impact of the feature map φ, and it does not involve any training process. The consequences on training (and generalization error) will be considered in Section 7.6.

7.4 Low WS of random features

We start by showing that the word sensitivity for the RF map in (7.1) is *low*, i.e., $\mathcal{S}_{\mathsf{RF}} = o(1)$.

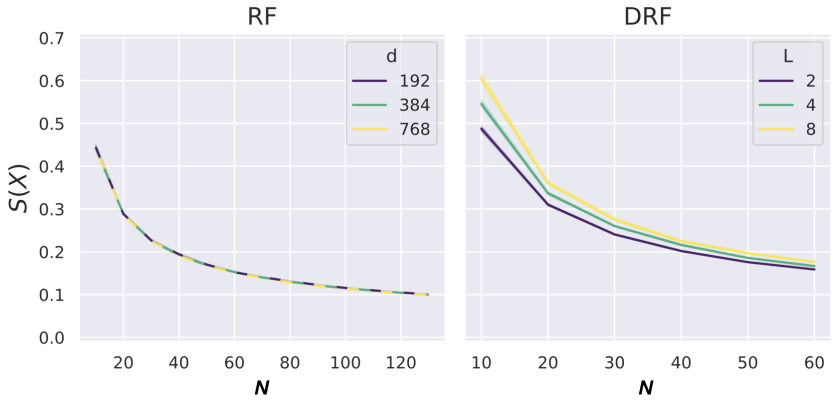


Figure 7.2: Numerical estimate of the WS for the RF (left) and DRF (right) map, with a ReLU activation function. We estimate $\mathcal{S}_\varphi$ looking for the perturbation $\Delta^* = \arg\sup_{\|\Delta\|_2 \le \sqrt{d}} \|\varphi(X^1(\Delta)) - \varphi(X)\|_2$, where we fix the first token for symmetry. We find Δ^* by optimizing our objective with constrained gradient ascent. For RF, we consider $d \in \{192, 384, 768\}$, as N increases (taking the first d dimensions of the embeddings of the first N tokens). For DRF, we repeat the experiment for different depths $L \in \{2, 4, 8\}$ and fixed $d = 768$, as N increases. As textual data X, we use the BERT-Base token embeddings of samples from the imdb dataset, after a pre-processing to adapt the dimensions and fulfill Assumption 7.1. We plot the average over 10 independent trials and the confidence band at 1 standard deviation. In the figure on the left, we intentionally dash the plotted lines to ease the visualization, as they overlap.

Theorem 7.2. *Let φ_{RF} be the random features map defined in* (7.1)*, where ϕ is Lipschitz and not identically* 0*. Let $X \in \mathbb{R}^{N\times d}$ be a generic input sample s.t. Assumption 7.1 holds, and assume $k = \Omega(D)$. Let $\mathcal{S}_{\mathsf{RF}}(X)$ denote the word sensitivity of φ_{RF} defined in* (7.4)*. Then, we have*

$$\mathcal{S}_{\mathsf{RF}}(X) = O\left(1/\sqrt{N}\right) = o(1), \tag{7.5}$$

with probability at least $1 - \exp(-cD)$ over V.

Theorem 7.2 shows that the RF model has a low word sensitivity, vanishing with the length of the context N. This means that, regardless of how any word is modified, the RF mapping is *not sensitive* to this modification, when the length of the context is large. The proof follows from an upper bound on the numerator of (7.4), due to the Lipschitz continuity of ϕ, and a concentration result on the norm at the denominator. The details are deferred to Appendix H.1.

Remark. Theorem 7.2 is readily extended to the case where the number of words m that can be modified in the context is $o(N)$. In fact, to achieve a sensitivity of constant order, a constant fraction of rows in X must be changed, i.e. $m = \Theta(N)$. This extension of Theorem 7.2 is also proved in Appendix H.1, and it further illustrates that a fully connected architecture is unable to capture the change of few (namely, $m = o(N)$) semantically relevant words in a long sentence.

We numerically validate this result in Figure 7.2 (left), where we estimate the value of $\mathcal{S}_{\mathsf{RF}}$ for different token embedding dimensions d, as the context length N increases. Clearly, $\mathcal{S}_{\mathsf{RF}}$ decreases as the context length increases, regardless of the embedding dimension d. In fact, the curves corresponding to different values of d (in different colors) basically coincide.

Deep Random Features (DRF). As an extension of Theorem 7.2, we show that the *word sensitivity is low* also for *deep* random features. We consider the DRF map

$$\varphi_{\text{DRF}}(X) := \phi(V_L\phi(V_{L-1}(...V_2\phi(V_1\,\text{flat}(X))...))), \tag{7.6}$$

where $\phi : \mathbb{R} \to \mathbb{R}$ is the component-wise non-linearity, k is the number of neurons at each layer, and $V_l \in \mathbb{R}^{k\times k}$ are the random weights at layer l, with $[V_1]_{i,j} \sim_{\text{i.i.d.}} \mathcal{N}(0, \beta/D)$ and $[V_l]_{i,j} \sim_{\text{i.i.d.}} \mathcal{N}(0, \beta/k)$ for $l > 1$. We set β according to He's initialization [He et al., 2015], which ensures

$$\mathbb{E}_{\rho\sim\mathcal{N}(0,\beta)}\left[\phi^2(\rho)\right] = 1, \tag{7.7}$$

as done in [Hanin, 2018] to avoid the problem of vanishing/exploding gradients in the analysis of a deep network.

Theorem 7.3. *Let φ_{DRF} be the deep random feature map defined in* (7.6)*, where ϕ is Lipschitz and β is chosen s.t.* (7.7) *holds. Let $X \in \mathbb{R}^{N\times d}$ be a generic input sample s.t. Assumption 7.1 holds, and assume $k = \Theta(D)$, $L = o(\log k)$. Let $\mathcal{S}_{\text{DRF}}(X)$ denote the word sensitivity of φ_{DRF} defined in* (7.4)*. Then, we have*

$$\mathcal{S}_{\text{DRF}} = O\left(\frac{e^{CL}}{\sqrt{N}}\right), \tag{7.8}$$

with probability at least $1 - \exp(-c\log^2 d)$ over $\{V_l\}_{l=1}^L$.

Theorem 7.3 shows that, as in the shallow case, the word sensitivity decreases with the length of the context N. The proof requires showing that $\|\varphi_{\text{DRF}}(X)\|_2$ concentrates to $\sqrt{k}$, which is achieved via (7.7). The strategy to bound the term $\|\varphi_{\text{DRF}}(X^i(\Delta)) - \varphi_{\text{DRF}}(X)\|_2$ is similar to that of the shallow case. The details are deferred to Appendix H.2.

The exponential dependence on L comes from our worst-case analysis of the Lipschitz constant of the model, and it does not fully exploit the independence between the V_i's. Thus, we expect the actual dependence of the WS on L to be milder. This is confirmed by the numerical results of Figure 7.2 (right), where we numerically estimate $\mathcal{S}_{\text{DRF}}$ for different depths L as the context length N increases. For all values of L, the word sensitivity quickly decreases with N.

7.5 High WS of random attention features

In contrast with random features, we show that the *word sensitivity is high* for the RAF map in (7.3), i.e., $\mathcal{S}_{\text{RAF}} = \Omega(1)$.

Theorem 7.4. *Let φ_{RAF} be the random attention features map defined in* (7.3)*. Let $X \in \mathbb{R}^{N\times d}$ be a generic input sample s.t. Assumption 7.1 holds, and assume $d/\log^4 d = \Omega(N)$. Let $\mathcal{S}_{\text{RAF}}(X)$ denote the word sensitivity of φ_{RAF} defined in* (7.4)*. Then, we have*

$$\mathcal{S}_{\text{RAF}}(X) = \Omega(1), \tag{7.9}$$

with probability at least $1 - \exp(-c\log^2 d)$ over W.

Theorem 7.4 shows that the RAF model has a high word sensitivity, regardless of the length of the context N, as long as $d/\log^4 d = \Omega(N)$. This requires a number of tokens N that grows slower than the embedding dimension d. Models such as BERT-Base or BERT-Large have an embedding dimension of 768 and 1024 [Devlin et al., 2019], which allows our results to hold for fairly large context lengths. In fact, we prove a stronger statement: for *any* index $i \in [N]$, there exists a perturbation Δ^* (possibly dependent on i) s.t. the RAF map changes significantly when evaluated in $X^i(\Delta^*)$. The proof of Theorem 7.4 is deferred to Appendix H.3, and a sketch follows.

Proof sketch. The argument is not constructive, and the difficulty in finding a closed-form solution for the perturbation Δ^* is due to the lack of assumptions on the sample X, which is entirely generic. We follow the steps below.

***Step 1**: Find a direction δ^* aligned with many words x_j's.* Using the probabilistic method, we prove the existence of a vector $\delta^* \in \mathbb{R}^d$, with $\|\delta^*\|_2 \leq \sqrt{d}$, s.t. its inner product with the tokens embeddings x_j's is large for a constant fraction of the words in the context. In particular, Lemma H.6 shows that $\left(x_j^\top \delta^*\right)^2 = \Omega(d^2/N)$ for at least $\Omega(N)$ indices $j \in [N]$.

***Step 2**: Exhibit two directions Δ_1^* and Δ_2^* both aligned with many words in the feature space $\{W^\top x_j\}_{j=1}^N$.* Exploiting the properties of the Gaussian attention features W, we deduce the existence of two vectors Δ_1^* and Δ_2^* that *(i)* are far from each other, and *(ii)* ensure a constant fraction of the entries of $XW\Delta_k^*/\sqrt{d}$, $k \in \{1,2\}$, to be large. In particular, by exploiting our assumption $d/\log^4 d = \Omega(N)$, Lemmas H.7 and H.8 show that $[XW\Delta_k^*/\sqrt{d}]_j = \Omega\left(\log^2 d\right)$ for at least $\Omega(N)$ indices $j \in [N]$, with $\|\Delta_1^* - \Delta_2^*\|_2 = \Omega(\sqrt{d})$.

***Step 3**: Show that the attention concentrates towards the perturbed word.* Recall that $s(X) := \mathrm{softmax}(XWX^\top/\sqrt{d})$ denotes the attention scores matrix. Then, Lemma H.9 proves that the attention scores $[s(X^i(\Delta_k^*))]_{j:}^\top$ are well approximated by the canonical basis vector e_i, for $k \in \{1,2\}$ and an $\Omega(N)$ number of rows $j \in [N]$. This intuitively means that a constant fraction of tokens x_j *moves all their attention* towards the i-th modified token $x_i + \Delta_k^*$. This step critically exploits the $\mathrm{softmax}$ function in the RAF map: if one entry in its argument is $\Omega\left(\log^2 d\right)$, then the attention scores concentrate as described above.

***Step 4**: Conclude with at least one perturbation between Δ_1^* and Δ_2^*.* Finally, exploiting the fact that $\|\Delta_1^* - \Delta_2^*\|_2 = \Omega(\sqrt{d})$, Lemma H.10 proves that at least one of them gives a Δ^* s.t. $\|\varphi_{\mathsf{RAF}}(X) - \varphi_{\mathsf{RAF}}(X^i(\Delta^*))\|_F = \Omega(\sqrt{dn})$. This, together with the upper bound $\|\varphi_{\mathsf{RAF}}(X)\|_F = \mathcal{O}(\sqrt{dn})$, concludes the argument. □

In Figure 7.3 (first plot), we estimate the value of $\mathcal{S}_{\mathsf{RAF}}$ for different token embedding dimensions d, as the context length N increases. In contrast with the random features map, even for large values of N, the WS remains larger than 1. In the same figure, in the second plot, we repeat the experiment for the ReLU-RAF map, which replaces the $\mathrm{softmax}$ with a ReLU activation. $\mathcal{S}_{\mathsf{ReLU\text{-}RAF}}(X)$ seems to decrease with the context length N, and has in general smaller values than $\mathcal{S}_{\mathsf{RAF}}(X)$. This highlights the importance of the $\mathrm{softmax}$ function, as discussed in *Step 3* of the proof sketch above. The condition $d/\log^4 d = \Omega(N)$ is required to allow *step 2* to go through. We believe this to be a reasonable assumption, as the maximum context length tends to be smaller than the embedding dimension. Popular examples include

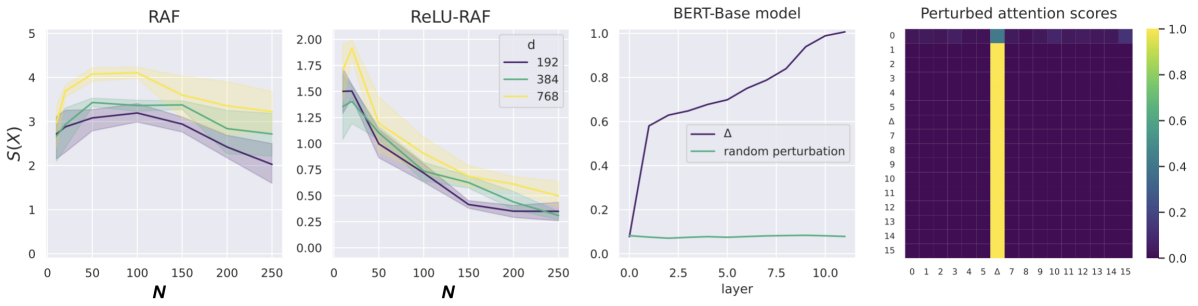


Figure 7.3: Numerical estimate of the WS for the RAF (first plot) and ReLU-RAF (second plot) map. The ReLU-RAF map is defined as the RAF one, but the $\mathrm{softmax}$ is replaced with a ReLU activation over the entries of $S(X)$, followed by a re-normalization to ensure that the attention scores sum up (on every row) to 1, as in the $\mathrm{softmax}$ case. We consider $d = 192, 384, 768$, as N increases. The rest of the setup is equivalent to the one described in Figure 7.2. In the third plot, we present the relative change of the pre-trained BERT-Base model layer embeddings, evaluated on the abstract of this chapter ($\sim$ 200 tokens), when the 42nd token is modified in embedding space with a vector Δ. The 0-th layer represents the input itself and, as a comparison, we report the results when the perturbation is chosen to be Gaussian noise with the same norm. In the fourth plot, we present the attention scores in the first head of the first layer of BERT-Base model, evaluated on the title of this chapter, when the 6th token is modified in embedding space with a vector Δ. In the third and fourth plots, Δ is chosen to be a perturbation that *attracts* all the attention on the perturbed key token, which follows the proof idea of Theorem 7.4.

BERT-Base ($N = 512$, $d = 768$), BERT-Large ($N = 512$, $d = 1024$), and the Llama-2 family ($N = 4096$, $d = 5120$).

Remark. The re-parameterization of the attention layer through the features W, which removes the dependence on queries, keys and values, does not substantially change the problem. In fact, *Step 2* of the argument uses that W acts as an approximate isometry on δ^*. This would also hold for the product of two independent Gaussian matrices $W_Q^\top W_K$. Similarly, introducing the independent Gaussian matrix W_V would not interfere with our conclusion in *Step 4*. Additionally, the results obtained on the RAF model seem to extend to realistic, pre-trained, transformer architectures. In fact, in the third plot of Figure 7.3, we provide a lower bound on the sensitivity of the BERT-Base architecture, evaluated on the abstract of this chapter, when the 42nd token is modified by a perturbation Δ. The blue line shows how after the first layer the embeddings are heavily modified, and how this change increases deeper in the architecture. In this case, the perturbation Δ follows from the proof idea of Theorem 7.4, as it is chosen to *move all the attention* towards x_i (see the fourth plot in Figure 7.3). To do so, we set Δ to be aligned with the right singular vector of $W_Q^\top W_K$ associated with the largest singular value. We do this on all the heads in the first layer separately, and then average and re-normalize.

7.6 Generalization on context modification

The study of the word sensitivity is motivated by understanding the capabilities of a model to learn to distinguish two contexts X and $X^i(\Delta)$ that only differ by a word. In fact, in practice, modifying a single row/word of X can lead to a significant change in the meaning,

see Figure 7.1. We formalize the problem in a supervised learning setting, and characterize whether a *generalized linear model* (GLM) induced by a feature map φ generalizes over a sample $(X^i(\Delta), y_\Delta)$, after being trained on (X, y). Crucially, the index $i \in [N]$ and the perturbation Δ are s.t. the label y_Δ is different from y (e.g., $y \in \{-1, +1\}$ and $y_\Delta = -y$), namely, perturbing the i-th word changes the meaning of the context.

Supervised learning with generalized linear models (GLMs). Let $(\mathcal{X}, \mathcal{Y})$ be a labelled training dataset, where $\mathcal{X} = (X_1, \dots, X_n)$ contains the training data $X_i \in \mathbb{R}^{N \times d}$ and $\mathcal{Y} = [y_1, \dots, y_n]^\top \in \{-1, 1\}^n$ the corresponding binary labels. The sample (X, y) does not belong to $(\mathcal{X}, \mathcal{Y})$, and it is introduced later in the training set. Let $\varphi : \mathbb{R}^{N \times d} \to \mathbb{R}^p$ be a feature map, and consider the GLM

$$f_\varphi(\cdot, \theta) = \varphi(\cdot)^\top \theta, \tag{7.10}$$

where $\theta \in \mathbb{R}^p$ are trainable parameters of the model. We define the feature matrix as $\Phi_\varphi := [\varphi(X_1), \dots, \varphi(X_n)]^\top \in \mathbb{R}^{n \times p}$, and focus on the quadratic loss

$$\mathcal{L}(\theta) := \frac{1}{n} \sum_{j=1}^{n} \left(\varphi(X_j)^\top \theta - y_j \right)^2. \tag{7.11}$$

Minimizing (7.11) with gradient descent gives (see equation (33) in [Bartlett et al., 2021])

$$\theta^* = \theta_0 + \Phi_\varphi^+ (\mathcal{Y} - \Phi_\varphi \theta_0), \tag{7.12}$$

where θ^* is the gradient descent solution, θ_0 is the initialization, and Φ_φ^+ is the Moore-Penrose inverse of Φ_φ.

Our goal is to establish whether the additional training on the sample (X, y) allows the model f_φ to generalize on $(X^i(\Delta), y_\Delta)$, with $y_\Delta = -y$. Thus, we do *not* focus on the case where $f_\varphi(X, \theta^*) = y$ and $f_\varphi(X^i(\Delta), \theta^*) = y_\Delta$, i.e., $f_\varphi(\cdot, \theta^*)$ already generalizes well on the two new samples. Instead, we look at how $f_\varphi(\cdot, \theta^*)$ *extrapolates* the information contained in the pair (X, y) to the perturbed sample $X^i(\Delta)$. This motivates the following assumption.

Assumption 7.5. *There exists a parameter $\gamma \in [0, 2)$ s.t.*

$$\left| f_\varphi(X^i(\Delta), \theta^*) - f_\varphi(X, \theta^*) \right| \leq \gamma. \tag{7.13}$$

The parameter γ captures the degree over which the trained model $f_\varphi(\cdot, \theta^*)$ can distinguish between X and $X^i(\Delta)$. If $\gamma = 0$, $f_\varphi(\cdot, \theta^*)$ does not recognize any difference between X and $X^i(\Delta)$; instead, if $f_\varphi(\cdot, \theta^*)$ correctly classifies the two samples X and $X^i(\Delta)$ (i.e., $f_\varphi(X, \theta^*) = y$ and $f_\varphi(X^i(\Delta), \theta^*) = y_\Delta = -y$), then $\gamma = 2$, which is beyond the scope of our analysis.

We consider training on (X, y) in the following two ways.

(a) Fine-tuning. First, we look at the model obtained after *fine-tuning* the solution θ^* defined in (7.12) over the new sample (X, y). This means that θ^* is the initialization of a gradient

descent algorithm trained on this single sample. As $\left(\varphi(X)^\top\right)^+ = \varphi(X)/\left\|\varphi(X)\right\|_2^2$, the fine-tuned solution is given by

$$\theta_f^* = \theta^* + \frac{\varphi(X)}{\|\varphi(X)\|_2^2}\left(y - \varphi(X)^\top\theta^*\right). \tag{7.14}$$

(b) Re-training. Second, we re-train the model from scratch, after adding the pair (X, y) to the training set. The new training set is denoted by $(\mathcal{X}_r, \mathcal{Y}_r)$, where $\mathcal{X}_r = (X_1, \ldots, X_n, X)$ contains the training data and $\mathcal{Y}_r = [y_1, \ldots, y_n, y]$ the binary labels. Thus, denoting by $\Phi_{\varphi,r} := [\varphi(X_1), \ldots, \varphi(X_n), \varphi(X)]^\top \in \mathbb{R}^{(n+1)\times p}$ the new feature matrix, the re-trained solution θ_r^* takes the form

$$\theta_r^* = \theta_0 + \Phi_{\varphi,r}^+(\mathcal{Y}_r - \Phi_{\varphi,r}\theta_0). \tag{7.15}$$

The quantity of interest is the *test error* on $(X^i(\Delta), y_\Delta)$:

$$\mathsf{Err}_\varphi(X^i(\Delta), \theta) := \left(f_\varphi(X^i(\Delta), \theta) - y_\Delta\right)^2, \tag{7.16}$$

where $\theta \in \{\theta_f^*, \theta_r^*\}$ is the vector of parameters obtained either after fine-tuning or re-training.

7.6.1 Random features do not generalize

By exploiting the low word sensitivity of random features, we show that both the fine-tuned and retrained solutions generalize poorly.

Theorem 7.6. *Let φ_{RF} be the random features map defined in* (7.1)*, with ϕ Lipschitz and not identically* 0*, and let $f_{\mathsf{RF}}(\cdot, \theta_f^*) = \varphi_{\mathsf{RF}}(\cdot)^\top\theta_f^*$ be the corresponding model fine-tuned on the sample (X, y), where $X \in \mathbb{R}^{N\times d}$ satisfies Assumption 7.1 and θ_f^* is given by* (7.14)*. Assume $k = \Omega(D)$, $|f_{\mathsf{RF}}(X, \theta^*)| = O(1)$, and that Assumption 7.5 holds with $\gamma \in [0, 2)$. Let $\mathsf{Err}_{\mathsf{RF}}(X^i(\Delta), \theta_f^*)$ be the test error of $f_{\mathsf{RF}}(\cdot, \theta_f^*)$ on $(X^i(\Delta), y_\Delta)$ as defined in* (7.16)*. Then, for any Δ s.t. $\|\Delta\|_2 \le \sqrt{d}$ and any $i \in [N]$, we have*

$$\mathsf{Err}_{\mathsf{RF}}(X^i(\Delta), \theta_f^*) > (2-\gamma)^2 - O\left(1/\sqrt{N}\right), \tag{7.17}$$

with probability at least $1 - \exp(-cD)$ over V.

Proof sketch. The idea is to use (7.14) to obtain that

$$f_{\mathsf{RF}}(X^i(\Delta), \theta_f^*) - f_{\mathsf{RF}}(X^i(\Delta), \theta^*) = \frac{\varphi_{\mathsf{RF}}(X^i(\Delta))^\top\varphi_{\mathsf{RF}}(X)}{\|\varphi_{\mathsf{RF}}(X)\|_2^2}\left(y - f_{\mathsf{RF}}(X, \theta^*)\right). \tag{7.18}$$

Next, we note that

$$\left|\frac{\varphi_{\mathsf{RF}}(X^i(\Delta))^\top\varphi_{\mathsf{RF}}(X)}{\|\varphi_{\mathsf{RF}}(X)\|_2^2} - 1\right| \le \mathcal{S}_{\mathsf{RF}}(X). \tag{7.19}$$

Theorem 7.2 gives that the RHS of (7.19) is small, which combined with (7.18) implies that $f_{\mathsf{RF}}(X^i(\Delta), \theta_f^*)$ is close to

$$y + f_{\mathsf{RF}}(X^i(\Delta), \theta^*) - f_{\mathsf{RF}}(X, \theta^*). \tag{7.20}$$

By upper bounding $f_{\mathsf{RF}}(X^i(\Delta), \theta^*) - f_{\mathsf{RF}}(X, \theta^*)$ via Assumption 7.5, we obtain that (7.20) cannot be far from y. This implies that $f_{\mathsf{RF}}(X^i(\Delta), \theta_f^*)$ cannot be close to the correct label $y_\Delta = -y$. The complete proof is in Appendix H.1. □

Theorem 7.7. *Let φ_{RF} be the random features map defined in* (7.1)*, with ϕ Lipschitz and non-linear, and let $f_{\mathsf{RF}}(\cdot, \theta_r^*) = \varphi_{\mathsf{RF}}(\cdot)^\top \theta_r^*$ be the corresponding model re-trained on the dataset $(\mathcal{X}_r, \mathcal{Y}_r)$ that contains the pair (X, y), thus with θ_r^* defined in* (7.15)*. Assume the training data to be sampled i.i.d. from a distribution $\mathcal{P}_X$ s.t. $\mathbb{E}_{X\sim\mathcal{P}_X}[X] = 0$, Assumption 7.1 holds, and the Lipschitz concentration property is satisfied. Let $n \log^3 n = o(k)$, $n \log^4 n = o(D^2)$ and $k = \Omega(D)$. Assume that $|f_{\mathsf{RF}}(X, \theta^*)| = O(1)$ and that Assumption 7.5 holds with $\gamma \in [0, 2)$. Let $\mathsf{Err}_{\mathsf{RF}}(X^i(\Delta), \theta_r^*)$ be the test error of $f_{\mathsf{RF}}(\cdot, \theta_r^*)$ on $(X^i(\Delta), y_\Delta)$ defined in* (7.16)*. Then, for any Δ s.t. $\|\Delta\|_2 \leq \sqrt{d}$ and any $i \in [N]$, we have*

$$\mathsf{Err}_{\mathsf{RF}}(X^i(\Delta), \theta_r^*) > (2 - \gamma)^2 - O\left(1/\sqrt{N}\right), \tag{7.21}$$

with probability at least $1 - \exp\left(-c \log^2 n\right)$ over $V, \mathcal{X}_r$.

Proof sketch. The idea is to leverage the stability analysis in [Bombari and Mondelli, 2024a], which gives

$$f_{\mathsf{RF}}(X^i(\Delta), \theta_r^*) - f_{\mathsf{RF}}(X^i(\Delta), \theta^*) = \mathcal{F}_{\mathsf{RF}}(X, X^i(\Delta)) \left(f_{\mathsf{RF}}(X, \theta_r^*) - f_{\mathsf{RF}}(X, \theta^*)\right), \tag{7.22}$$

where

$$\mathcal{F}_{\mathsf{RF}}(X, X^i(\Delta)) := \frac{\varphi_{\mathsf{RF}}(X^i(\Delta))^\top P_\Phi^\perp \varphi_{\mathsf{RF}}(X)}{\left\|P_\Phi^\perp \varphi_{\mathsf{RF}}(X)\right\|_2^2} \tag{7.23}$$

is the *feature alignment* between X and $X^i(\Delta)$ induced by φ_{RF} and P_Φ the projector over $\mathrm{Span}\{\varphi_{\mathsf{RF}}(X_1), \dots, \varphi_{\mathsf{RF}}(X_N)\}$. After some manipulations, we have

$$\left|\mathcal{F}_{\mathsf{RF}}(X, X^i(\Delta)) - 1\right| \leq \mathcal{S}_{\mathsf{RF}}(X) \frac{\|\varphi_{\mathsf{RF}}(X)\|_2}{\sqrt{\lambda_{\min}\left(K_{\mathsf{RF},r}\right)}}, \tag{7.24}$$

where $K_{\mathsf{RF},r} := \Phi_{\mathsf{RF},r} \Phi_{\mathsf{RF},r}^\top$ is the kernel of the model. A lower bound on its smallest eigenvalue $\lambda_{\min}\left(K_{\mathsf{RF},r}\right)$ follows from the fact that the kernel is well-conditioned (see Lemma H.2), which crucially relies on the assumptions on the data (i.i.d. and Lipschitz concentrated) and the scalings $n \log^3 n = o(k)$, $n \log^4 n = o(D^2)$. As the word sensitivity $\mathcal{S}_{\mathsf{RF}}(X)$ is upper bounded by Theorem 7.2, from (7.24) we conclude that $\mathcal{F}_{\mathsf{RF}}(X, X^i(\Delta))$ is close to 1.

Since $K_{\mathsf{RF},r}$ is invertible, the re-trained model $f_{\mathsf{RF}}(\cdot, \theta_r^*)$ interpolates the dataset $(\mathcal{X}_r, \mathcal{Y}_r)$, giving that $f(X, \theta_r^*) = y$. Thus, as $\mathcal{F}_{\mathsf{RF}}(X, X^i(\Delta)) \approx 1$, $f_{\mathsf{RF}}(X^i(\Delta), \theta_r^*)$ is close to (7.20), and we conclude from the same argument used for Theorem 7.6. The complete proof is in Appendix H.1. □

In a nutshell, by exploiting the low word sensitivity of random features, Theorems 7.6-7.7 show that, after either fine-tuning or re-training, the model does not learn to "separate" the predictions on the samples X and $X^i(\Delta)$. As a consequence, the test error is lower bounded by $(2 - \gamma)^2$. In fact, γ is the distance between the predictions on X and $X^i(\Delta)$ before fine-tuning/re-training (see (7.13)), and the ground-truth labels have distance 2 ($y_\Delta, y \in \{-1, 1\}$ and $y_\Delta = -y$).

While Theorem 7.6 does not require distributional assumptions on the data, Theorem 7.7 considers i.i.d. training data, satisfying Lipschitz concentration. This property corresponds to having well-behaved tails, and it is common in the related theoretical literature [Bubeck and Sellke, 2021, Nguyen et al., 2021b, Bombari et al., 2022b], see Appendix B for the formal definition and a discussion.

We remark that Assumption 7.5 requires the model $f_{\mathsf{RF}}(\cdot, \theta^*)$ to give a similar output when evaluated on the two new samples X and $X^i(\Delta)$. Thus, we are asking if the model generalizes on $X^i(\Delta)$ *only* from the additional training on X. Now, one could design an adversarial Δ s.t. $f(X^i(\Delta), \theta_r^*)$ and $f(X, \theta_r^*)$ are different from each other (so that $f_{\mathsf{RF}}(X^i(\Delta), \theta_r^*) = y_\Delta$ while $f_{\mathsf{RF}}(X, \theta_r^*) = y$), by exploiting the adversarial vulnerability of random features [Dohmatob and Bietti, 2022, Dohmatob, 2022, Bombari et al., 2023]. However, if we restrict the possible Δ's to those that satisfy Assumption 7.5, Theorems 7.6 and 7.7 prove that such adversarial patch cannot be found. We finally note that, when the context length N is comparable or larger than the number of training samples n, the model becomes adversarially robust to any token modification and Assumption 7.5 automatically holds, see Appendix H.4 for details.

7.6.2 Random attention features can generalize

Next, the behavior of random features is contrasted with that of random attention features. Let us consider the RAF model $f_{\mathsf{RAF}}(\cdot, \theta) := \mathrm{flat}(\varphi_{\mathsf{RAF}}(\cdot))^\top \theta$, where $\varphi_{\mathsf{RAF}}(\cdot)$ is defined in (7.3). Theorem 7.4 proves that the word sensitivity of $\varphi_{\mathsf{RAF}}(\cdot)$ is large. This suggests that the RAF model is capable of extrapolating the information contained in (X, y) to correctly classify the perturbed sample $X^i(\Delta)$. While proving a rigorous statement on a fine-tuned/re-trained RAF model remains challenging, we provide experimental evidence of this generalization capability.

Figure 7.4 (first row) shows that, after fine-tuning on (X, y), $f_{\mathsf{RAF}}(X^i(\Delta), \theta_f^*)$ can be close to the perturbed label $y_\Delta = -y$, even if the model before fine-tuning was unable to distinguish between X and $X^i(\Delta)$. Specifically, the two central sub-plots consider the RAF model for two values of the context length $N \in \{40, 120\}$: here, the loss on the perturbed sample can be close to 0, even when the parameter γ in (7.13) is close to 0, i.e., X and $X^i(\Delta)$ were indistinguishable before fine-tuning; in general, the test error $\mathrm{Err}_{\mathsf{RAF}}(X^i(\Delta), \theta_f^*)$ is often smaller than the lower bound of $(2-\gamma)^2$ (dashed black line), which holds for random features.

The left sub-plot considers the RF model for $N = 40$: here, the loss on the perturbed sample is close to 0 only if the model before fine-tuning was already able to perfectly distinguish between X and $X^i(\Delta)$, i.e., γ is not far from 2; in general, the test error $\mathrm{Err}_{\mathsf{RF}}(X^i(\Delta), \theta_f^*)$ always respects the lower bound of $(2-\gamma)^2$ proved in Theorem 7.6. Finally, the right sub-plot considers the ReLU-RAF model (which replaces the $\mathrm{softmax}$ with a ReLU activation, as described at the end of Section 7.5) for $N = 120$: here, even if the lower bound of Theorem 7.6 is often violated, the model still cannot reach small error unless γ is large, i.e., X and $X^i(\Delta)$ could be distinguished already before fine-tuning. This confirms the impact of the $\mathrm{softmax}$ on the capability of attention layers to understand the context. Analogous results hold when models are re-trained (instead of being fine-tuned), as reported in the second row of the same figure.

In a nutshell, our results show that, when a RAF model is not able to distinguish two points with opposite labels (that only differ in one word), fine-tuning or retraining on one of these points allows the loss on also the other point to decrease. In contrast, this is not the case

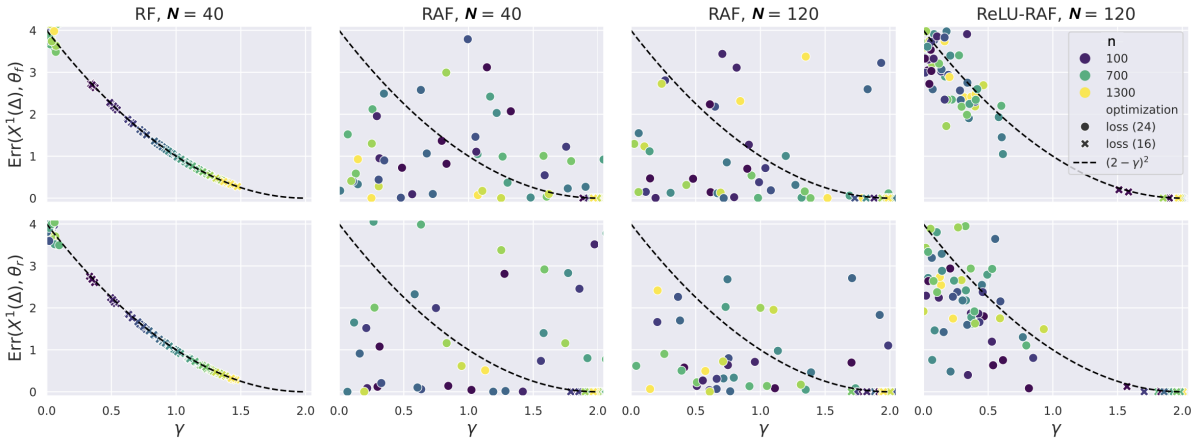


Figure 7.4: Test error (as defined in (7.16) taking $i = 1$) for the RF (left subplot), RAF (two central sub-plots) and ReLU-RAF (right subplot) maps, as a function of the smallest γ s.t. Assumption 7.5 holds. The first (resp. second) row considers the fine-tuned solution θ_f^* (resp. re-trained solution θ_r^*). Every sub-plot has a fixed embedding dimension $d = 768$, and context length $N \in \{40, 120\}$, taking the first N token embeddings for each sample. Different colors correspond to a different number of training samples $n \in \{100, 700, 1300\}$. Every point in the scatter-plots is an independent simulation where (X, y) and $(\mathcal{X}, \mathcal{Y})$ are the BERT-Base embeddings of a random subset of the imdb dataset (after pre-processing to fulfill Assumption 7.1). Circular markers correspond to obtaining Δ via gradient descent optimization of the losses in (7.25); cross markers correspond to minimizing directly the test error in (7.16).

for the RF model, where the loss on the second point can be lower bounded according to Theorems 7.6 and 7.7. This is shown in Figure 7.4, where most of the points for the RAF model are below the dashed line representing the lower bound for the RF model.

7.6.3 Experimental details

The experiments of Figure 7.4 are performed on multiple independent trials, for different choices of the training data. We report in cross markers the results obtained by choosing Δ after optimizing the test error in (7.16) via gradient descent. While this approach directly minimizes the metric of interest, it results in low test error only when γ is rather large, regardless of the model taken into account (RF, RAF, or ReLU-RAF). In contrast, optimizing a different loss controls the value of γ, while still achieving small error for RAF (and, to a smaller extent, ReLU-RAF). We report in circular markers the results obtained by minimizing the following two losses (respectively, for fine-tuning and re-training):

$$\ell_{\theta_f^*}(\Delta) := \left(\frac{\varphi_{\mathsf{RAF}}(X^i(\Delta))^\top \varphi_{\mathsf{RAF}}(X)}{\|\varphi_{\mathsf{RAF}}(X)\|_2^2} + 1 \right)^2, \qquad \ell_{\theta_r^*}(\Delta) := \left(\mathcal{F}_{\mathsf{RAF}}(X, X^i(\Delta)) + 1 \right)^2. \tag{7.25}$$

This choice is suggested by (7.18) and (7.22) which, after assuming for simplicity that $f_{\mathsf{RAF}}(X^i(\Delta), \theta^*) = f_{\mathsf{RAF}}(X, \theta^*) = 0$, can be re-written as

$$f_{\mathsf{RAF}}(X^i(\Delta), \theta_f^*) = \frac{\varphi_{\mathsf{RAF}}(X^i(\Delta))^\top \varphi_{\mathsf{RAF}}(X)}{\|\varphi_{\mathsf{RAF}}(X)\|_2^2}\, y, \qquad f_{\mathsf{RAF}}(X^i(\Delta), \theta_r^*) = \mathcal{F}_{\mathsf{RAF}}(X, X^i(\Delta))\, y. \tag{7.26}$$

Achieving small error means that the LHS of (7.26) is close to $-y$, which corresponds to making the losses in (7.25) small. Further numerical results and comparisons between different optimization algorithms for finding Δ are in Appendix H.5.

7.7 Conclusions

This chapter provides a formal characterization of the fundamental difference between fully connected and attention layers. To do so, we consider the prototypical setting of random features and study the *word sensitivity*, which captures how the output of a map changes after perturbing a single row/word of the input. On the one hand, the sensitivity of standard random features decreases with the context length and, in order to obtain to a significant change in the output of the map, a constant fraction of the words needs to be perturbed. On the other hand, the sensitivity of random attention features is large, regardless of the context length, thus indicating the suitability of attention layers for NLP tasks. These bounds on the word sensitivity translate into formal negative generalization results for random features, which are contrasted by positive empirical evidence of generalization for the attention layer.

Our analysis allows the perturbations to be any (bounded) vector in the embedding space. Taking the tokenization process (and, hence, the discrete nature of the textual samples) into account offers an exciting avenue for future work.

CHAPTER 8

Privacy for Free in the Overparameterized Regime

Differentially private gradient descent (DP-GD) is a popular algorithm to train deep learning models with provable guarantees on the privacy of the training data. In the last decade, the problem of understanding its performance cost with respect to standard GD has received remarkable attention from the research community, which formally derived upper bounds on the excess population risk $\mathcal{R}_{\mathrm{P}}$ in different learning settings. However, existing bounds typically degrade with over-parameterization, *i.e.*, as the number of parameters p gets larger than the number of training samples n – a regime which is ubiquitous in current deep-learning practice. As a result, the lack of theoretical insights leaves practitioners without clear guidance, leading some to reduce the effective number of trainable parameters to improve performance, while others use larger models to achieve better results through scale. In this chapter, we show that in the popular *random features* model with quadratic loss, for *any* sufficiently large p, privacy can be obtained for free, *i.e.*, $|\mathcal{R}_{\mathrm{P}}| = o(1)$, not only when the privacy parameter ε has constant order, but also in the strongly private setting $\varepsilon = o(1)$. This challenges the common wisdom that over-parameterization inherently hinders performance in private learning.

This chapter is based on the work published in PNAS 2025: *Privacy for free in the overparameterized regime*.

8.1 Introduction

Deep learning models have been shown to be vulnerable to various attacks directed to retrieve information about the training dataset [Fredrikson et al., 2015, Carlini et al., 2019, 2021, Haim et al., 2022]. This vulnerability is a concern when sensitive, private data is included in the training set of a learning pipeline. To allow developers and model providers to still utilize these data, *differential privacy* (DP) [Dwork et al., 2006] consolidated as the golden standard for privacy in learning algorithms. In fact, this framework comes with algorithmic methods [Shokri and Shmatikov, 2015, Abadi et al., 2016] that provide formal protection guarantees for each individual sample in the training set, which becomes safeguarded (up to some level) by any adversary with access to the trained model and the rest of the dataset.

Neural networks can be trained in a differentially private fashion via algorithms such as DP (stochastic) gradient descent (DP-GD) [Abadi et al., 2016]. This method involves minimizing

the training loss with additional "tweaks" to guarantee protection, which typically boil down to *(i) clipping* the per-sample gradients before averaging, *(ii)* perturbing the parameters updates with *random noise*, and *(iii)* limiting the number of training iterations with *early stopping* (see Algorithm 8.1). While these adjustments provide privacy guarantees, they often come with a performance cost with respect to standard (stochastic) GD algorithms [Tramer and Boneh, 2021, Kurakin et al., 2022]. The impact on performance tends to worsen as the privacy requirements become stronger, as this requires greater perturbations to the gradients. Thus, private training of deep learning models involves carefully tuning additional hyper-parameters, e.g., the clipping constant, the magnitude of the noise, and the number of training iterations, in order to reach the best performance for a given privacy requirement. Notably, this comes at a remarkable computational cost, also considering the higher training times and memory loads that DP optimization has with respect to standard GD algorithms, mainly due to the computation of the per-sample gradients [De et al., 2022].

The challenging problem of optimizing neural networks with an assigned privacy guarantee has motivated a thriving field of research proposing novel architectures and training algorithms [Tramer and Boneh, 2021, Papernot et al., 2021, De et al., 2022]. Concurrently, theoretical studies have emerged with the scope of better understanding and quantifying the *privacy-utility* trade-offs in different learning settings. Here, *privacy* is often defined via the pair of parameters (ε, δ): the impact of a single data point on the output of the algorithm is controlled by ε with probability $1-\delta$, see Definition 8.1 for the formal introduction of (ε, δ)-DP. In order to provide meaningful protection, practitioners pick constant-order values of ε, *e.g.*, $\varepsilon \in \{1, 2, 4, 8\}$, and $\delta < 1/n$, where n is the number of training samples [Bassily et al., 2019]. The *utility*, in this context, is typically measured as the degradation in generalization performance of the DP-trained solution $\theta^p \in \mathbb{R}^p$ compared to a non-private baseline $\theta^* \in \mathbb{R}^p$, where p denotes the number of parameters of the model. In particular, considering the standard supervised setting and denoting by $(x, y) \sim \mathcal{P}_{XY}$ an input-label pair distributed accordingly to the data distribution $\mathcal{P}_{XY}$, the *excess population risk* is defined as

$$\mathcal{R}_{\mathsf{P}} = \mathbb{E}_{(x,y)\sim\mathcal{P}_{XY}}\left[\ell(x, y, \theta^p)\right] - \mathbb{E}_{(x,y)\sim\mathcal{P}_{XY}}\left[\ell(x, y, \theta^*)\right], \tag{8.1}$$

where $\ell(x, y, \theta)$ is the loss over the sample (x, y) of the model evaluated in θ. Intuitively, the excess population risk worsens with more stringent privacy requirements on θ^p (*i.e.*, with smaller values of ε and δ), and a rich line of work spanning over a decade has investigated this trade-off in various settings, see e.g. [Kifer et al., 2012, Jain and Thakurta, 2014, Bassily et al., 2014, 2019, Song et al., 2021b, Varshney et al., 2022, Brown et al., 2024b].

Despite this flurry of research, existing results are unable to address the over-parameterized regime, *i.e.*, $p = \Omega(n)$. In fact, in this regime, taking ε of constant order (*e.g.*, $\varepsilon = 1$) typically makes the upper bound on the excess population risk to read (at best) $\mathcal{R}_{\mathsf{P}} = O(1)$, which is vacuous as the performance of a trivial model that constantly outputs zero is of the same order. For additional details, we refer to the discussion on related works in Section 8.2. This result is sometimes understood via the qualitative argument that the noise introduced at each iteration by DP-GD increases with the dimension of the parameter space. More specifically, if each entry of $\theta \in \mathbb{R}^p$ is perturbed with standard Gaussian noise, the ℓ_2 norm of the perturbation scales as $\sqrt{p}$, which suggests that the privacy-utility trade-off worsens as p grows [Yu et al., 2021a, Mehta et al., 2022]. This motivated a line of research that focuses on DP algorithms acting over lower dimensional subspaces, both in theory [Zhou et al., 2021b, Asi et al., 2021] and in practice [Yu et al., 2021b, Golatkar et al., 2022].

However, reducing the dimensionality of architectures is in stark contrast with the current trend of increasingly large networks in deep learning. In the non-private setting, the apparent contradiction between excellent generalization of over-parameterized models and classical bias-variance trade-off has been the subject of intense investigation, which has unveiled phenomena such as benign overfitting [Bartlett et al., 2021, 2020] and double descent [Belkin et al., 2019a, Mei and Montanari, 2022].

Recently, larger models have also shown benefits on down-stream tasks requiring private fine-tuning [Li et al., 2022b, Yu et al., 2022], which motivated theoretical studies that provide refined privacy-utility trade-offs in this specific setting [Wang et al., 2024, Li et al., 2022a]. Perhaps surprisingly, the recent work by De et al. [2022] gives evidence of the benefits of scale even in the absence of public pre-training data, as the generalization performance is shown to improve with model size on datasets such as CIFAR-10 and ImageNet, following an accurate hyper-parameter search.

In Figure 8.1, we investigate the interplay between privacy and over-parameterization in a simpler and more controllable setting: training a 2-layer, fully-connected ReLU network on MNIST ($n = 6 \cdot 10^4$) with DP-GD (see Algorithm 8.1). We vary the network width, spanning both the under-parameterized ($p \sim 10^4$) and over-parameterized ($p \sim 10^7$) regime, questioning if the algorithm suffers as the number of parameters grows larger. The plot shows that this is not the case: the test accuracy initially increases until the network is wide enough ($p \sim 10^5$), and then plateaus without showing any decreasing trend in the right side of the figure.

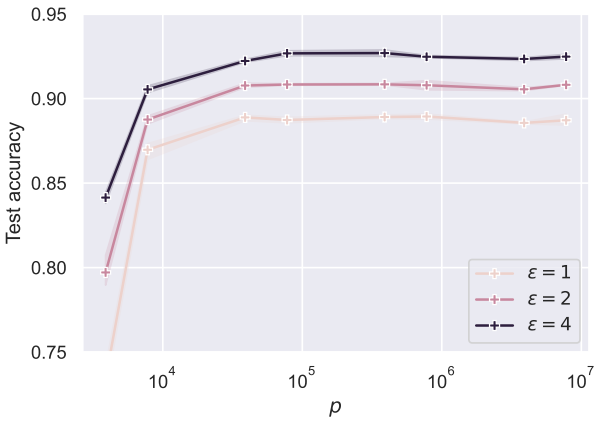


Figure 8.1: Test accuracy of DP-GD on MNIST for a 2-layer, fully connected ReLU network, as a function of the number of parameters p with fixed $n = 50000$. Further details on the experimental setting can be found in Section 8.5.

Then, are larger models truly disadvantageous when doing private optimization? How can the result in the simple experiment in Figure 8.1 be reconciled with existing theoretical work on privacy-utility trade-offs in the over-parameterized regime? How can one solve the disagreement between the experimental evidence in De et al. [2022] and the widespread idea that more parameters hurt the performance of DP-GD? These questions call for a principled understanding of private training in deep learning. This in particular amounts to *(i)* establishing meaningful generalization guarantees (i.e., $\mathcal{R}_{\mathsf{P}}$ significantly smaller than trivial guessing at test time) for over-parameterized models, and *(ii)* characterizing how the hyperparameters introduced by DP-GD (clipping constant, noise magnitude, and number of training iterations) affect performance.

8.1.1 Main contribution

In this work, we provide privacy-utility guarantees $\mathcal{R}_{\mathsf{P}} = o(1)$ under over-parameterization – not only when ε has constant order, but also in the strongly private setting $\varepsilon = o(1)$. We frame this result as achieving *privacy for free* in the over-parameterized regime.

We consider a family of machine learning models where the number of parameters p is a free variable that can be significantly larger than the number of samples n and the input dimension d, just as in the experimental setup of Figure 8.1. This excludes, for example, linear regression, where $p = d$ [Varshney et al., 2022, Liu et al., 2023b, Brown et al., 2024b]. Specifically, we focus on the widely studied *random features* (RF) model [Rahimi and Recht, 2007] with quadratic loss, which takes the form

$$\ell(x, y, \theta) = \left(f_{\mathsf{RF}}(x, \theta) - y\right)^2, \qquad f_{\mathsf{RF}}(x, \theta) = \phi(Vx)^\top \theta, \tag{8.2}$$

where f_{RF} is a generalized linear model, $\phi : \mathbb{R} \to \mathbb{R}$ a non-linearity applied component-wise to the vector $Vx \in \mathbb{R}^p$, and $V \in \mathbb{R}^{p \times d}$ a random weight matrix. The RF model can be regarded as a 2-layer fully-connected neural network, where only the second layer θ is trained and the hidden weights V are frozen at initialization. Its appeal comes from the fact that it is simple enough to be analytically tractable and, at the same time, rich enough to exhibit properties occurring in more complex deep learning models [Bartlett et al., 2021, Mei and Montanari, 2022].

The DP-trained solution θ^p is obtained via DP-GD as reported in Algorithm 8.1. Differently from [Abadi et al., 2016], we do not perform stochastic batching of the data when aggregating the gradients. The cost of privacy typically refers to the performance gap between an algorithm that enforces (ε, δ)-DP and one that does not. Thus, we set the non-private baseline θ^* in (8.1) to be the solution obtained by (non-private) GD. In the RF model, the test error of this baseline has been characterized precisely in a recent line of work [Mei and Montanari, 2022, Mei et al., 2021, Hu et al., 2024] and, for a class of data distributions, θ^* also corresponds to the Bayes-optimal solution given by $\arg\min_\theta \mathbb{E}_{(x,y)\sim\mathcal{P}_{XY}}\left[\ell(x, y, \theta)\right]$. At this point, we can present an informal version of our result (formally stated as Theorem 8.9).

Theorem (main result – informal)**.** Consider the RF model in (8.2) with input dimension d and number of features p. Let n be the number of training samples and $\mathcal{R}_{\mathsf{P}}$ be defined according to (8.1), where θ^* is the solution of GD and θ^p is the (ε, δ)-differentially private solution of DP-GD (Algorithm 8.1). Then, for all sufficiently over-parameterized models, under some technical conditions, the following holds with high probability

$$|\mathcal{R}_{\mathsf{P}}| = \tilde{O}\left(\frac{d}{n\varepsilon} + \sqrt{\frac{d}{n}} + \sqrt{\frac{n}{d^{3/2}}}\right) = o(1). \tag{8.3}$$

In words, in the regime $d \ll n \ll d^{3/2}$ – considered e.g. in [Mei et al., 2021, Bombari et al., 2023], see Assumption 8.11 for details – we show that $|\mathcal{R}_{\mathsf{P}}| = o(1)$ as long as $\varepsilon \gg d/n$. In fact, when $d \ll n \ll d^{3/2}$ and $\varepsilon \gg d/n$, the three terms $\frac{d}{n\varepsilon}$, $\sqrt{\frac{d}{n}}$ and $\sqrt{\frac{n}{d^{3/2}}}$ appearing in the RHS of (8.3) become $o(1)$. We make two observations: *(i)* as $d \ll n$, our result guarantees vanishing excess population risk, even with a strong privacy requirement $\varepsilon = o(1)$; *(ii)* the bound in (8.3) does not depend on the number of parameters p and Assumption 8.11 only requires a lower bound on p, which allows for arbitrarily large over-parameterization in the model. As typical in the related literature, the dependence of $\mathcal{R}_{\mathsf{P}}$ on δ is only logarithmic and, therefore, it is neglected in the notation $\tilde{O}(\cdot)$ that hides poly-logarithmic factors in δ and n.

Achieving *privacy for free* might seem surprising, so we now comment on the intuition behind the result. In the regime $n \gg d$, there is a surplus of samples that can be used to learn privately. In fact, in our model, the test error of (non-private) GD plateaus when n is between d and $d^{3/2}$ [Mei et al., 2021, Hu et al., 2024], hence $\Theta(d)$ samples are enough to achieve

utility, with the remaining ones playing the role of achieving *privacy*. The plateau in the test loss of GD has been shown for other kernel ridge regression models [Ghorbani et al., 2021], and in RF it is exhibited when $d^l \ll n \ll d^{l+1}$ for any $l \in \mathbb{N}$, as long as $p \gg n$, see Figure 1 of Hu et al. [2024]. We expect that DP-GD catches up with the performance of GD in any of these plateaus. However, as n approaches d^{l+1}, the test loss of GD sharply decreases and it is unclear whether DP-GD has the same rate of improvement. This suggests that our result could be extended to the regime $d^l \ll n \ll d^{l+1}$ with $p \gg n$. In this chapter, we focus on $d \ll n \ll d^{3/2}$ and $p = \Omega(n^2)$ due to the additional challenges in the analysis of clipping, see the discussion after Lemma 8.14 in Section 8.6.

Figure 8.2 displays the intuition of having a surplus of samples. There, we plot the test accuracy of a fixed model as the number of training samples n increases, both for GD and DP-GD, keeping p fixed and much larger than n. In the left side of the figure, the performance of θ^* sharply increases with n, while the private algorithms have very low utility. Moving towards the right side of the plot, the performance of GD tends to saturate, thus approaching the regime where there is a surplus of samples required to learn the task. Here, the extra samples are used to achieve privacy: the utility of DP-GD increases, eventually approaching the non-private baseline.

Our main result proves that, in the RF model, this heuristic picture holds for any sufficiently large degree of over-parameterization. The key technical hurdle is to formalize these intuitions. Specifically, while Mei et al. [2021], Hu et al. [2024] establish the asymptotic test error of θ^* at convergence, we need a *non-asymptotic* control (in terms of n, d, p) on the *whole DP-GD trajectory* to understand the impact of clipping and early stopping. This in turn leads to a completely different proof strategy: we first frame DP-GD as the Euler-Maruyama discretization scheme of a stochastic differential equation (SDE); then, exploiting a number of concentration arguments (Dudley, Sudakov-Fernique, and Borell-TIS inequalities), we establish the size of a set containing the full trajectory of the SDE (with high probability), which allows to set the clipping constant; finally, we study the DP-trained solution θ^p via an Ornstein-Uhlenbeck process, which allows to set the early stopping. Remarkably, our approach provides insights on the problem of hyper-parameter optimization. In fact, as a byproduct of our analysis, we establish the scaling of key hyper-parameters (clipping constant, noise magnitude and number of training iterations) in terms of the learning setup (input dimension d, number of parameters p) for random features.

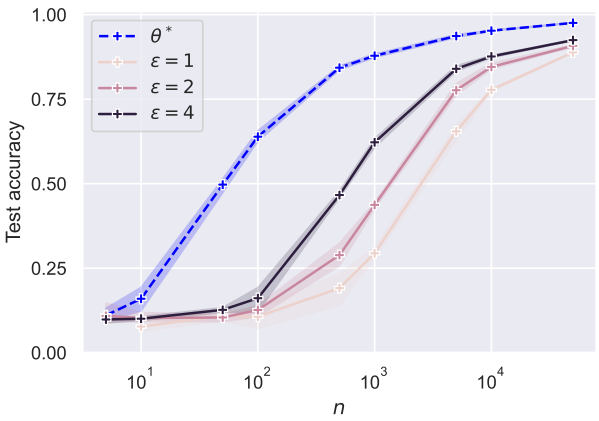


Figure 8.2: Test accuracy of DP-GD on MNIST for a 2-layer, fully connected ReLU network, as a function of the number of training samples n with fixed hidden layer width $= 1000$. Further details on the experimental setting can be found in Section 8.5.

8.2 Related work

Differentially private empirical risk minimization is the problem of minimizing the empirical risk $\mathcal{L}(\theta) = \sum_{i=1}^{n} \ell(x_i, y_i, \theta)/n$ while also ensuring privacy guarantees, where (x_i, y_i) represents an input-label pair in the training set and ℓ is a loss function. Different algorithms have been proposed to tackle this problem, and the performance is measured through the excess risks of their solution θ^p with respect to the baseline θ^*. In particular, we can define the *excess empirical risk* as the difference in the training losses $\mathcal{R}_\mathsf{E} = \mathcal{L}(\theta^p) - \mathcal{L}(\theta^*)$, and the *excess population risk* $\mathcal{R}_\mathsf{P}$ as the difference in the test losses, see (8.1). These quantities capture the privacy-utility trade-off of the DP algorithm, as excess risks worsen with an increase in the privacy guarantees (*i.e.*, with smaller values of the privacy parameter ε). The seminal works by Chaudhuri and Monteleoni [2008], Chaudhuri et al. [2011] quantitatively investigate this trade-off for strongly convex losses in the context of ε-DP (obtained by setting $\delta = 0$), both for output and objective perturbation methods.

Constrained optimization. Kifer et al. [2012] extend the analysis of the objective perturbation by Chaudhuri et al. [2011], providing a bound on the empirical excess risk of $\mathcal{R}_\mathsf{E} = \tilde{O}(\sqrt{p}/(n\varepsilon))$ for constrained, strongly convex (ε, δ)-DP optimization (see their Theorem 4). The subsequent work by Bassily et al. [2014] focuses on a variant of DP-GD previously studied by Williams and Mcsherry [2010], and it bounds the excess empirical risk as $\mathcal{R}_\mathsf{E} = \beta L \tilde{O}(\sqrt{p}/(n\varepsilon))$ (see their Theorem 2.4.2).[1] Here, β is the diameter of the (bounded) optimization domain $\mathcal{B}$, L is the Lipschitz constant of the loss with respect to the parameters of the model, and the baseline is defined as the training loss minimizer $\theta^* = \arg\min_{\theta \in \mathcal{B}} \mathcal{L}(\theta)$. We note that, in this setting, clipping is not necessary as the loss is Lipschitz. By relying on the algorithmic stability analysis in Shalev-Shwartz et al. [2009], Bassily et al. [2014] also provide an excess population risk bound of the form $\mathcal{R}_\mathsf{P} = \tilde{O}(p^{1/4}/\sqrt{n\varepsilon})$ (see their Theorem F.2.2), where the baseline is the Bayes optimal solution $\theta^* = \arg\min_{\theta \in \mathcal{B}} \mathbb{E}_{(x,y)\sim\mathcal{P}_{XY}} [\ell(x, y, \theta)]$. Using again algorithmic stability [Hardt et al., 2016], Bassily et al. [2019] later improve this bound to $\mathbb{E}[\mathcal{R}_\mathsf{P}] = \beta L \tilde{O}(1/\sqrt{n} + \sqrt{p}/(n\varepsilon))$, where the expectation is taken with respect to the randomness of the training dataset and algorithm.

Unconstrained optimization. In the setting of (convex) unconstrained optimization, Jain and Thakurta [2014] aim to remove the dependence of p in excess risk bounds for (ε, δ)-DP, which was present in earlier work by Chaudhuri et al. [2011]. They focus on generalized linear models (GLMs, *i.e.*, $\ell(x, y, \theta) = \ell(\varphi(x)^\top \theta, y)$ where φ is the feature map), assume L-Lipschitz loss with a strongly convex regularization term, and bound the excess population risk with respect to the Bayes optimal solution θ^* as $\mathcal{R}_\mathsf{P} = \tilde{O}(L\|\theta^*\|_2/\sqrt{n\varepsilon})$. The more recent work by Song et al. [2021b] also considers unconstrained convex Lipschitz GLMs, without the assumption on the strong convexity. Using the same non-private baseline θ^* and denoting by M the projector on the column space of $\mathbb{E}_{x\sim\mathcal{P}_X}[\varphi(x)\varphi(x)^\top]$, the previous bound is improved by $\mathbb{E}[\mathcal{R}_\mathsf{P}] = L\,\|M\theta^*\|_2\, \tilde{O}\left(1/\sqrt{n} + \sqrt{\min\left(\operatorname{rank}(M), n\right)/(\varepsilon n)}\right)$ (see their Theorem 3.2). Li et al. [2022a] consider Lipschitz losses with ℓ_2 regularization and introduce the notion of restricted Lipschitz continuity, which gives dimension-free bounds. However, in the setting of GLMs, this approach leads to the same bounds as in [Song et al., 2021b]. A similar approach is taken by Ma et al. [2022] that remove the dependence on p at the cost of an additional factor

[1] The definition of empirical excess risk in Bassily et al. [2014] is not re-normalized by n. This makes the bounds of Theorem 2.4.2 therein larger by a factor n, which is removed here for the sake of comparison with other works.

$\mathrm{tr}(\tilde{H})$, where $\tilde{H} \succeq \sup_\theta \nabla^2_\theta \mathbb{E}_{(x,y)\sim\mathcal{P}_{XY}}[\ell(x,y,\theta)]$. The work by Arora et al. [2022] considers GLMs with a loss that is not necessarily Lipschitz, and it also recovers the result of Song et al. [2021b] in the Lipschitz setting with $\varepsilon = \Omega(1)$.

Importantly, even in the prototypical case of an RF model, existing bounds do not cover the over-parameterized regime. We now provide a detailed discussion explaining why this is the case. First, note that the RF model in (8.2) has a non-Lipschitz (quadratic) loss, which does not allow a direct application of most previous bounds. Thus, to ensure a fair comparison, one can estimate $\|\theta^*\|_2$ and evaluate the Lipschitz constant of the training loss restricted to a bounded set $\mathcal{B}$ with radius $\beta = \|\theta^*\|_2$. This provides the scaling of the effective Lipschitz constant of the model, as if the optimization was bounded to the set $\mathcal{B}$. From our results in the later sections, it can be shown that $\|\theta^*\|_2 = \Theta(\sqrt{n/p})$, which in turn gives $\|\theta^*\|_2 \sup_{\theta\in\mathcal{B}} \|\nabla_\theta \ell(x_i, y_i, \theta)\|_2 = \Theta(n)$. This makes the bound by Jain and Thakurta [2014] read $\mathcal{R}_\mathsf{P} = \tilde{O}(\sqrt{n}/\varepsilon)$, which is vacuous for $\varepsilon = \Theta(1)$. A similar discussion holds for the bound by Bassily et al. [2019]. Furthermore, even by assuming that the loss function in (8.2) is Lipschitz, the result improves only by a factor $\sqrt{n}$, which gives $\mathcal{R}_\mathsf{P} = \tilde{O}(1/\varepsilon)$ and, therefore, a trivial bound when $\varepsilon = \Theta(1)$. Similar considerations apply to the bound by Song et al. [2021b] (and, thus, to the ones by Li et al. [2022a], Arora et al. [2022]), since in this setting we have that $\|M\theta^*\|_2 = \|\theta^*\|_2$ and $\mathrm{rank}(M) \geq n$. As concerns the result by Ma et al. [2022], it can be verified that $\mathrm{tr}(\tilde{H}) \geq \mathbb{E}_x\left[\|\varphi(x)\|_2^2\right] = \Theta(p)$, which reintroduces the dependence on the number of parameters p, preventing an improvement upon Jain and Thakurta [2014]. Finally, while Arora et al. [2022] give bounds for quadratic losses, the reasoning resembles the one above on the effective Lipschitz constant in $\mathcal{B}$ and it does not lead to an improvement with respect to the Lipschitz case. The detailed calculation of the quantities mentioned in this paragraph is deferred to Appendix I.1.

Linear regression. This model reads $\ell(x, y, \theta) = (x^\top\theta - y)^2$, where $y = x^\top\theta^* + \sigma$, θ^* is the ground-truth solution and σ independent label noise. Varshney et al. [2022] assume that each point is sampled from a standard Gaussian distribution and obtain a sample complexity of d/ε via gradient descent with adaptive clipping [Andrew et al., 2021]. This improves over [Wang, 2018, Cai et al., 2021, Milionis et al., 2022], and it matches the result of earlier work by Liu et al. [2022], which however proposes a computationally inefficient method. Liu et al. [2023b] relax the assumption on the data (sub-Weibull concentration) and improve the dependence of the sample complexity on the condition number of the covariance matrix. Brown et al. [2024b] remove the need for adaptive clipping, focusing on a setting where the clipping constant is large enough so that clipping will most likely never happen.

8.3 Differentially Private Gradient Descent

The formal definition of DP builds on the notion of *adjacent datasets*. In our supervised learning setting, a dataset D' is said to be adjacent to a dataset D if they differ by only one sample.

Definition 8.1 ((ε,δ)-DP [Dwork et al., 2006])**.** *A randomized algorithm satisfies (ε,δ)-differential privacy if for any pair of adjacent datasets D, D', and for any subset of the parameters space $S \subseteq \mathbb{R}^p$, we have*

$$\mathbb{P}\left(\mathcal{A}(D) \in S\right) \leq e^\varepsilon \mathbb{P}\left(\mathcal{A}(D') \in S\right) + \delta. \tag{8.4}$$

Algorithm 8.1: DP-GD

Input: Number of iterations T, learning rate η, clipping constant C_{clip}, noise σ, initialization θ_0.

for $t \in [T]$ **do**

  Compute the gradient $g(x_i, y_i, \theta_{t-1}) \leftarrow \nabla_\theta \ell(x_i, y_i, \theta_{t-1})$;

  Clip the gradients $g_{C_{\text{clip}}}(x_i, y_i, \theta_{t-1}) \leftarrow g(x_i, y_i, \theta_{t-1}) / \max\left(1, \frac{\|g_t(x_i, y_i, \theta_{t-1})\|_2}{C_{\text{clip}}}\right)$;

  Aggregate the gradients $g_{\theta_{t-1}} \leftarrow \frac{1}{n} \sum_i g_{C_{\text{clip}}}(x_i, y_i, \theta_{t-1})$;

  Update the model parameters adding noise $\theta_t = \theta_{t-1} - \eta g_{\theta_{t-1}} + \sqrt{\eta} \frac{2C_{\text{clip}}}{n} \sigma \mathcal{N}(0, I_p)$;

end

Output: Model parameters θ_T.

Here, the probability is with respect to the randomness induced by the algorithm, and the inequality has to hold uniformly on all the adjacent datasets D and D'.

Different methods have been proposed to train machine learning models while enforcing (ε, δ)-DP. An approach is to rely on the scheme of *output perturbation*, which adds noise to the output of empirical risk minimization [Dwork et al., 2006, Chaudhuri et al., 2011, Kifer et al., 2012]. In deep learning, however, it is not easy to characterize the final parameters of the model, and therefore to estimate the minimal amount of noise required to guarantee protection. For this reason, a popular choice relies on DP-GD algorithms, which involve perturbing the individual updates during training time, thus providing privacy guarantees based on the size of the perturbation and the number of iterations [Song et al., 2013, Bassily et al., 2014, Abadi et al., 2016]. In this work, we consider Algorithm 8.1, which can be regarded as a variant of the well-established method by Abadi et al. [2016] without stochastic batching. Its privacy guarantees are reported below.

Proposition 8.2. *For any $\delta \in (0, 1)$, $\varepsilon \in (0, 8 \log(1/\delta))$, if we set*

$$\sigma \geq \sqrt{\eta T} \frac{\sqrt{8 \log(1/\delta)}}{\varepsilon}, \tag{8.5}$$

then Algorithm 8.1 is (ε, δ)-differentially private.

The proof follows the strategy of Abadi et al. [2016] and is based on the moment accountant method. As our formulation is slightly different, we provide the complete argument in Appendix I.2.

While existing work tackles the problem of understanding utility through stability arguments [Bassily et al., 2019, Song et al., 2021b], we propose a novel approach, which focuses on the implicit bias of DP-GD. The idea is to resort to a continuous process defined by Algorithm 8.1 in the limit of small learning rates $\eta \to 0$. This has proven effective in the non-private setting, where the *gradient flow* equation helped understanding the properties of GD in over-parameterized models [Mei and Montanari, 2022, Mei et al., 2021, Hu et al., 2024]. However, two challenges arise in the analysis of DP-GD: *(i)* the *clipping* of the per-sample gradients, and *(ii)* the injection of *noise*. To overcome the former, we define an auxiliary *clipped loss* $\mathcal{L}_{C_{\text{clip}}}(\theta)$ that incorporates clipping. As for the latter, we consider the stochastic differential equation (SDE) obtained by adding a Wiener process to the gradient flow. This

motivates the scaling of the standard deviation of the injected noise as $\sqrt{\eta}$, as e.g. considered by Wang et al. [2019].

More specifically, similarly to Song et al. [2021b], we note that the clipping step in Algorithm 8.1 can be reformulated as the optimization of the surrogate loss $\ell_{i,C_{\text{clip}}}(\varphi(x_i)^\top\theta - y_i)$, whose derivative for the i-th training sample reads

$$\ell'_{i,C_{\text{clip}}}(z) = \ell'(z)\min\left(1, \frac{C_{\text{clip}}}{|\ell'(z)|\,\|\varphi(x_i)\|_2}\right). \tag{8.6}$$

This is formalized by the result below, whose proof is deferred to Appendix I.2.

Proposition 8.3. *For any $\theta \in \mathbb{R}^p$ and any clipping constant $C_{\text{clip}} > 0$, we have that $g_{C_{\text{clip}}}(x_i, y_i, \theta) = \nabla_\theta \ell_{i,C_{\text{clip}}}\left(\varphi(x_i)^\top\theta_{t-1} - y_i\right)$, where $g_{C_{\text{clip}}}(x_i, y_i, \theta)$ is defined in the clipping step of Algorithm 8.1.*

In other words, running Algorithm 8.1 is equivalent to running the same algorithm, without the clipping step, on the *clipped loss* $\mathcal{L}_{C_{\text{clip}}}(\theta) = \frac{1}{n}\sum_{i=1}^n \ell_{i,C_{\text{clip}}}(\varphi(x_i)^\top\theta - y_i)$. Hence, we can write the t-th iteration of Algorithm 8.1 as a noisy GD update on $\mathcal{L}_{C_{\text{clip}}}(\theta)$, *i.e.*,

$$\theta_t - \theta_{t-1} = -\eta\nabla\mathcal{L}_{C_{\text{clip}}}(\theta_{t-1}) + \sqrt{\eta}\frac{2C_{\text{clip}}}{n}\sigma\mathcal{N}(0, I_p). \tag{8.7}$$

This update rule corresponds to the Euler-Maruyama discretization scheme of the SDE

$$\mathrm{d}\Theta(t) = -\nabla\mathcal{L}_{C_{\text{clip}}}(\Theta(t))\mathrm{d}t + \frac{2C_{\text{clip}}}{n}\sigma\,\mathrm{d}B(t) = -\nabla\mathcal{L}_{C_{\text{clip}}}(\Theta(t))\mathrm{d}t + \Sigma\,\mathrm{d}B(t), \tag{8.8}$$

with discretization η (see Section 10.2 of Kloeden and Platen [2011]). Here, $B(t)$ represents the standard p-dimensional Wiener process, we introduce the shorthand $\Sigma = 2C_{\text{clip}}\sigma/n$, and we set the initial condition of (8.8) to correspond to the initialization of Algorithm 8.1, *i.e.* $\Theta(0) = \theta_0$. The strong convergence of the Euler-Maruyama method guarantees that, for any $\tau = \eta T$, the solution θ_T of Algorithm 8.1 approaches the solution $\Theta(\tau)$ of the SDE in (8.8) as the learning rate η gets smaller. We note that previous work [Paquette et al., 2024a] has considered a similar SDE to analyze the effects of stochastic batching, separating the dynamics in a gradient flow plus a Wiener process. Thus, the approach developed here could prove useful also to study DP-SGD, after incorporating an additional independent Wiener process in (8.8). Importantly, privacy guarantees analogous to those of Proposition 8.2 hold for $\Theta(\tau)$, as formalized in the result below (the proof is deferred to Appendix I.2).

Proposition 8.4. *Let $\ell : \mathbb{R} \to \mathbb{R}$ be a differentiable function with Lipschitz-continuous derivative. Then, for any $\delta \in (0, 1)$, $\varepsilon \in (0, 8\log(1/\delta))$, if we set*

$$\Sigma \geq \frac{2C_{\text{clip}}\sqrt{\tau}}{n}\frac{\sqrt{8\log(1/\delta)}}{\varepsilon}, \tag{8.9}$$

the solution $\Theta(\tau)$ of the SDE (8.8) at time τ is (ε, δ)-differentially private.

Armed with Proposition 8.4, the rest of the chapter focuses on the analysis of the DP-trained solution $\Theta(\tau)$. In fact, by strong convergence, the privacy-utility trade-offs of θ_T approach the ones of $\Theta(\tau)$, for sufficiently small learning rates. As a non-private baseline, we consider the solution of the gradient flow equation

$$\mathrm{d}\hat{\theta}(t) = -\nabla\mathcal{L}(\hat{\theta}(t))\mathrm{d}t, \qquad \theta^* = \lim_{t\to+\infty}\hat{\theta}(t), \tag{8.10}$$

where $\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(\varphi(x_i)^\top \theta - y_i)$ is the original training loss (see Section 5.1 of Bartlett et al. [2021] for more details on the convergence). Note that (8.10) resembles (8.8), without the effects of clipping and the introduction of noise, and θ^* is defined at convergence, as it does not require early stopping.

8.4 Main result

For simplicity, we set the initialization $\theta_0 = 0$, although similar results can be obtained for a broad class of initializations. The random features matrix $V \in \mathbb{R}^{p\times d}$ is a normalized Gaussian matrix, *i.e.*, $V_{i,j} \sim_{\text{i.i.d.}} \mathcal{N}(0, 1/d)$, and the data distribution is scaled as follows.

Assumption 8.5 (Data distribution)**.** *The training samples $\{(x_1, y_1), \dots, (x_n, y_n)\}$ are n i.i.d. samples from the joint distribution $\mathcal{P}_{XY}$, such that the marginal distribution $\mathcal{P}_X$ satisfies the following properties:*

1. *$x \sim \mathcal{P}_X$ is sub-Gaussian, with $\|x\|_{\psi_2} = O(1)$.*
2. *The data $x \sim \mathcal{P}_X$ are normalized such that $\|x\|_2 = \sqrt{d}$.*
3. *$\lambda_{\min}\left(\mathbb{E}_{x\sim\mathcal{P}_X}\left[xx^\top\right]\right) = \Omega(1)$,* i.e.*, the second-moment matrix of the data is well-conditioned.*

Furthermore, we assume all the labels $(y_1, \dots, y_n)$ to be bounded.

The first assumption requires the data to have well-behaved tails. It is commonly used in the literature [Liu et al., 2023b], and it covers a number of important cases such as data respecting Lipschitz concentration [Bubeck and Sellke, 2021] (*e.g.*, standard Gaussian [Varshney et al., 2022, Brown et al., 2024b] or uniform on the sphere [Mei and Montanari, 2022]). The second assumption (and the boundedness of the labels) can be achieved via data pre-processing. The third assumption is also required in related work to achieve the best rates in linear regression [Varshney et al., 2022, Liu et al., 2023b]. The scaling of input data and random features V guarantees that the pre-activations of the model (*i.e.*, the entries of Vx) are of constant order. We then process the entries of Vx via the following family of activation functions.

Assumption 8.6 (Activation function)**.** *The activation function $\phi : \mathbb{R} \to \mathbb{R}$ is a non-linear, Lipschitz continuous function such that $\mu_0 = \mu_2 = 0$ and $\mu_1 \neq 0$, where μ_k denotes the k-th Hermite coefficient of ϕ.*

This choice is motivated by theoretical convenience, and it covers a wide family of activations, including all odd ones (*e.g.*, $\tanh$). We believe that our result can be extended to a more general setting, as the one in Mei et al. [2021], at the cost of a more involved analysis.

Assumption 8.7 (Parameter scaling)**.**

$$n = O(\sqrt{p}), \qquad \log n = \Theta(\log p), \qquad n = \omega\left(d \log^2 d\right), \qquad n = o\left(\frac{d^{3/2}}{\log^3 d}\right). \quad (8.11)$$

We consider an over-parameterized setting in which the number of parameters p is at least of order n^2. To guarantee that the RF model interpolates the data, it suffices that $p \gg n$ (see

e.g. [Mei et al., 2021, Wang and Zhu, 2023]), and we expect our result to hold under this milder assumption. This would however need a different argument, specifically for Lemma 8.14. The requirement $\log n = \Theta(\log p)$ is mild and it could be relaxed at the expenses of a poly-logarithmic dependence on p in our final result in Theorem 8.9. Finally, we focus on the regime $d \ll n \ll d^{3/2}$, which corresponds to standard datasets, such as CIFAR-10 ($n = 5 \cdot 10^4$, $d \approx 3 \cdot 10^3$), or ImageNet as considered in [Zhang et al., 2017b] ($n \approx 1.3 \cdot 10^6$, $d \approx 9 \cdot 10^4$).

Assumption 8.8 (Privacy budget)**.**

$$\delta \in (0,1), \qquad \varepsilon \in (0, 8\log(1/\delta)), \qquad \frac{\varepsilon}{\sqrt{\log(1/\delta)}} = \omega\left(\frac{d\log^5 n}{n}\right). \tag{8.12}$$

The first two conditions on δ, ε are standard in differential privacy. The lower bound on ε coming from the third condition still allows for strong privacy regimes with $\varepsilon = o(1)$, as $n \gg d$ from Assumption 8.7. To state our main result, we set the hyper-parameters in the SDE (8.8) as

$$\tau = \frac{d\log^2 n}{p}, \qquad C_{\mathsf{clip}} = \sqrt{p}\log^2 n, \qquad \Sigma = \frac{2C_{\mathsf{clip}}\sqrt{\tau}}{n}\frac{\sqrt{8\log(1/\delta)}}{\varepsilon}. \tag{8.13}$$

Theorem 8.9. *Consider the RF model in* (8.2) *with input dimension d and number of features p. Let n be the number of training samples and $\mathcal{R}_{\mathsf{P}}$ be defined in* (8.1)*, where θ^* is given by* (8.10) *and θ^p is the solution $\Theta(\tau)$ of the SDE* (8.8) *at time τ, with hyper-parameters set as in* (8.13)*. Let Assumptions 8.5, 8.6, 8.7, 8.8 hold. Then, we have that θ^p is (ε, δ)-differentially private, and that*

$$\begin{aligned} |\mathcal{R}_P| &= O\left(\frac{d}{n\varepsilon}\,\log^5 n\sqrt{\log(1/\delta)} + \sqrt{\frac{d}{n}} + \sqrt{\frac{n\log^3 d}{d^{3/2}}}\right) \\ &= \tilde{O}\left(\frac{d}{n\varepsilon} + \sqrt{\frac{d}{n}} + \sqrt{\frac{n}{d^{3/2}}}\right) = o(1), \end{aligned} \tag{8.14}$$

with probability at least $1 - 2\exp\left(-c\log^2 n\right)$, where c is an absolute constant.

The strong convergence of the Euler-Maruyama discretization scheme guarantees that, for sufficiently small learning rates η, the solution θ_T of Algorithm 8.1 also satisfies (8.14), which establishes the utility of DP-GD.

8.5 Numerical experiments

Theorem 8.9 proves that, in the RF model, over-parameterization is not inherently detrimental in private learning. To verify this numerically, we plot in the first panel of Figure 8.3 the test loss of a family of RF models trained on a synthetic dataset via DP-GD, as the number of parameters p increases. As a comparison, we also report the performance achieved by (non-private) GD, which provides the baseline θ^*. While the test loss of θ^* displays the typical double-descent curve [Belkin et al., 2019a, Hastie et al., 2022], with the expected peak at the interpolation threshold ($p = n$), the performance of θ^p steadily improves and, as p increases, it plateaus to a value close to the corresponding test loss of θ^*. This is in agreement with Theorem 8.9, which predicts that, for a sufficiently over-parameterized model, there is a small

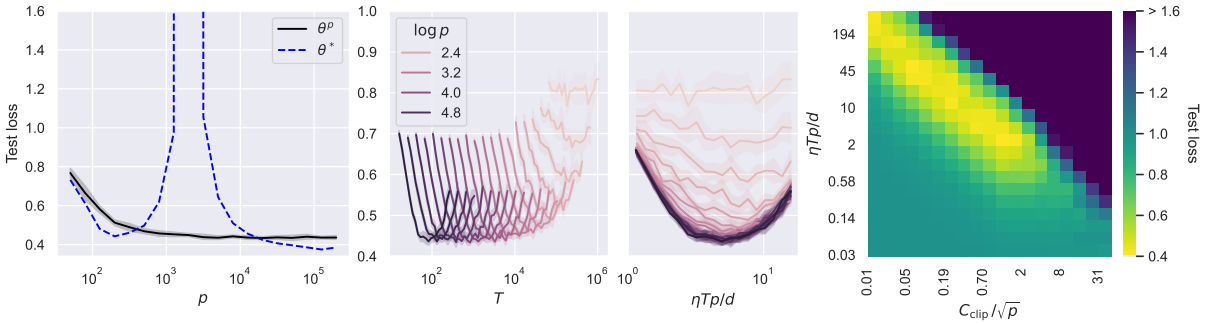


Figure 8.3: Experiments on RF models with $\tanh$ activation, and synthetic data sampled from a standard Gaussian distribution with $d = 100$. The learning task is given by $y = \text{sign}(u^\top x)$, where $u \in \mathbb{R}^d$ is a fixed vector sampled from the unit sphere, and we consider a fixed number of training samples $n = 2000$. θ^p is the solution of Algorithm 8.1 with $\varepsilon = 4$, $\delta = 1/n$, and θ^* is the solution of GD, both with small enough learning rate η. *First panel:* test losses of θ^p and θ^* for different number of parameters p. *Second panel:* test loss for θ^p as a function of the number of training iterations T. *Third panel:* Same plot as in the second panel, with the x-axis set to be $\eta T p/d$. *Fourth panel:* test loss of θ^p for a fixed $p = 40000$, as a function of the hyper-parameters (C_{clip}, T).

performance gap between θ^p and θ^*. Furthermore, the lack of an interpolation peak in the test loss of DP-GD points to the regularization offered by this algorithm and it resembles the effect of a ridge penalty, see [Raskutti et al., 2014] for a connection between ridge and early stopping and also the discussion after Lemma 8.13 in Section 8.6.

In the first panel of Figure 8.3, the hyper-parameters in DP-GD are chosen to maximize performance, and the scaling of such optimal hyper-parameters is considered in the other panels. Specifically, in the second plot, we take $C_{\text{clip}} = 0.5\sqrt{p}$ and report the test error as a function of the number of iterations T, setting the noise according to Proposition 8.2 to guarantee the desired privacy budget. As suggested by (8.13), the optimal T minimizing the test error decreases with p. Then, as $\eta T = \tau$ scales proportionally to d/p in (8.13), we rescale the plot putting $\eta T p/d$ on the x-axis of the third panel. The curves now collapse onto each other, confirming the accuracy of the proposed scaling. This enables us to transfer the hyper-parameters from small RF models (where brute-force optimization is affordable) to larger models (where training is more time and resources intensive). Finally, the heat-map in the rightmost panel displays the results of a full hyper-parameter grid search over (C_{clip}, T) for a fixed p and ε. We distinguish 4 regions in hyper-parameters space: in the *(1) top-right* we have very low utility due to the large noise required by the high number of training iterations T and large clipping constants

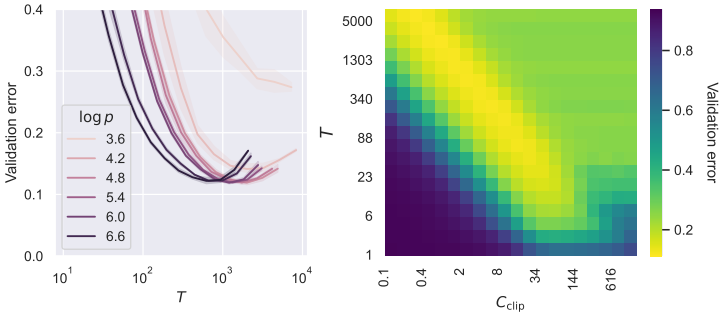


Figure 8.4: Experiments on a family of 2-layer fully-connected ReLU networks trained with cross-entropy loss on the MNIST classification task ($d = 768$ and we fix $n = 50000$), with privacy budget $\varepsilon = 1$, $\delta = 1/n$. *First panel:* validation error as a function of the number of training iterations T. *Second panel:* validation error for a hidden-layer width of 1000 ($p \sim 10^6$), as a function of the hyper-parameters (C_{clip}, T).

C_{clip}. In the *(2) bottom-left* the test loss is close to 1 as at initialization, since the model does not have the time to learn due to low T and small C_{clip}. In the *(3) bottom-right* we have larger C_{clip} which could allow for faster convergence. However, as C_{clip} becomes larger than the typical per-sample gradient size ($C_{\text{clip}} \gg \sqrt{p}$), the overly-pessimistic injection of noise ultimately undermines utility. In the *(4) top-left* we have *high* T and *small* C_{clip}, which lead to low test loss. At the center of the panel, we have $C_{\text{clip}} \sim \sqrt{p}$ and $\eta T \sim d/p$ as in (8.13), which also lead to low test loss. Moving towards the upper-left part of the plot there is no decrease in performance, as the slower rate of convergence (smaller C_{clip} and thus smaller gradients) is compensated by larger T, keeping fixed the overall amount of injected noise. Thus, the model has enough time to learn (differently from the bottom-left), while the added noise does not catastrophically degrade utility (differently from the top-right). This behaviour is in line with earlier empirical work [Kurakin et al., 2022, Li et al., 2022b, De et al., 2022] noting that wide ranges of C_{clip} result in optimal performance.

A similar picture emerges from the training of 2-layer neural networks with DP-GD and cross-entropy loss on MNIST. Figure 8.1 already showed that the performance of θ^p steadily improves and then plateaus as p increases, thus demonstrating that DP-GD benefits from larger models, contrary to the view that more parameters hinder privacy. Figure 8.4 then considers the scaling of the hyper-parameters (which are chosen to maximize performance in Figure 8.1). Specifically, in the first panel, we set $C_{\text{clip}} = 1$ and report the validation error as a function of T; in the second panel, we present the full hyper-parameter grid search over (C_{clip}, T). The results are qualitatively similar to those presented in Figure 8.3 for random features: on the left, the optimal T decreases with p; on the right, the grid exhibits the 4 regions discussed above. The implementation of the experiments is publicly available at the GitHub repository `https://github.com/simone-bombari/privacy-for-free`.

8.6 Proof of Theorem 8.9

Notation. All logarithms are in the natural basis. Random variables are defined on different probability spaces, which describe the randomness in the data, the random features and the private mechanisms. All complexity notations are understood for sufficiently large data size n, input dimension d and number of parameters p. We indicate with C and c absolute, strictly positive, numerical constants, that do not depend on the scalings of the problem and whose value may change from line to line. Given a positive number s, $[s]$ denotes the set of positive numbers from 1 to s. Given a vector v, $\|v\|_2$ denotes its Euclidean norm. Given a matrix A, $\|A\|_{\text{op}}$ denotes its operator norm, $\|A\|_F$ its Frobenius norm and A^+ its Moore-Penrose inverse. We use $X \in \mathbb{R}^{n\times d}$ to denote the data matrix (containing the i-th element of the training set $x_i \in \mathbb{R}^d$ in its i-th row), $\Phi \in \mathbb{R}^{n\times p}$ to denote the feature matrix (containing $\varphi(x_i) := \phi(Vx) \in \mathbb{R}^p$ in its i-th row) and $K := \Phi\Phi^\top \in \mathbb{R}^{n\times n}$ to denote the kernel associated with the feature map.

8.6.1 Technical outline

To circumvent the difficulty in explicitly solving the SDE in (8.8) (*i.e.*, in characterizing the probability density function of $\Theta(\tau)$), we instead consider the SDE

$$\mathrm{d}\hat{\Theta}(t) = -\nabla\mathcal{L}(\hat{\Theta}(t))\mathrm{d}t + \Sigma\,\mathrm{d}B(t) = -\frac{2\Phi^\top}{n}\left(\Phi\hat{\Theta}(t) - Y\right)\mathrm{d}t + \Sigma\,\mathrm{d}B(t), \tag{8.15}$$

where $\mathcal{L}(\theta)$ is the original (quadratic) training loss and $B(t)$ is the same standard Wiener process as in (8.8). The solution of the SDE in (8.15) is a multi-dimensional Ornstein–Uhlenbeck (OU) process which admits a closed form (see, e.g., Section 4.4.4 in Gardiner [1985]). Let us then define

$$\mathcal{C} := \left\{\theta \quad \text{s.t.} \quad \|\nabla_\theta \ell(\varphi(x_i)^\top \theta - y_i)\|_2 < C_{\text{clip}} \quad \text{for all } i \in [n]\right\}, \tag{8.16}$$

which corresponds to the subset of the parameters space where *clipping does not happen*, *i.e.*, where $\mathcal{L}_{C_{\text{clip}}}(\theta) = \mathcal{L}(\theta)$. If the full path of the process $\Theta(t)$ happens in this region (*i.e.* $\Theta(t) \in \mathcal{C}$ for all $t \in [0, \tau]$), then $\Theta(\tau) = \hat{\Theta}(\tau)$. This corresponds to the event

$$\hat{\Theta}(t) \in \mathcal{C}, \text{ for all } t \in [0, \tau], \tag{8.17}$$

which is easier to control, as $\hat{\Theta}(t)$ is an OU process. The first part of our proof consists in showing that, for our choice of the hyper-parameters in (8.13), this event happens with high probability. To do so, we consider the decomposition

$$\hat{\Theta}(t) = \mathbb{E}_B\left[\hat{\Theta}(t)\right] + \tilde{\Theta}(t) = \hat{\theta}(t) + \tilde{\Theta}(t), \tag{8.18}$$

where we introduce the notation $\tilde{\Theta}(t) := \hat{\Theta}(t) - \mathbb{E}_B[\hat{\Theta}(t)]$ and use that the expectation of an OU process corresponds to the gradient flow $\mathbb{E}_B[\hat{\Theta}(t)] = \hat{\theta}(t)$ defined in (8.10). Then, $\mathcal{C}$ can be characterized as

$$\begin{aligned} \mathcal{C} &:= \left\{\theta \quad \text{s.t.} \quad \left\|\nabla_\theta \ell(\varphi(x_i)^\top \theta - y_i)\right\|_2 < C_{\text{clip}} \quad \text{for all } i \in [n]\right\} \\ &= \left\{\theta \quad \text{s.t.} \quad \left|\varphi(x_i)^\top \theta - y_i\right| < \frac{C_{\text{clip}}}{2\left\|\varphi(x_i)\right\|_2} \quad \text{for all } i \in [n]\right\}, \end{aligned} \tag{8.19}$$

which implies that the probability of the event in (8.17) is lower bounded by the probability of the event

$$\left|\varphi(x_i)^\top \tilde{\Theta}(t)\right| + \left|\varphi(x_i)^\top \hat{\theta}(t) - y_i\right| \leq \frac{C_{\text{clip}}}{2\left\|\varphi(x_i)\right\|_2}, \quad \text{for all } i \in [n], \quad \text{for all } t \in [0, \tau]. \tag{8.20}$$

As we set $C_{\text{clip}} = \sqrt{p}\log^2 n$ (see (8.13)) and $\|\varphi(x_i)\|_2 = \Theta(\sqrt{p})$ with high probability (see the argument in (I.147) and (I.148)), we therefore want to show that, for all $i \in [n]$ and for all $t \in [0, \tau]$, $|\varphi(x_i)^\top \tilde{\Theta}(t)| = o(\log^2 n)$ and $|\varphi(x_i)^\top \hat{\theta}(t) - y_i| = o(\log^2 n)$ with high probability. The term $|\varphi(x_i)^\top \hat{\theta}(t) - y_i|$ is handled by the lemma below.

Lemma 8.10. *Let Assumptions 8.5, 8.6 and 8.7 hold. Then, we have that, jointly for all* $i \in [n]$,

$$\sup_{t \in [0,\tau]} \left|\varphi(x_i)^\top \hat{\theta}(t) - y_i\right| = O\left(\log n\right), \tag{8.21}$$

with probability at least $1 - 2\exp\left(-c\log^2 n\right)$ *over* V *and* X*, where* c *is an absolute constant.*

We now briefly sketch the argument to obtain (8.21), deferring the full proof to Section 8.6.2. As the initialization $\theta_0 = 0$, the solution of the gradient flow in (8.10) takes the form $\hat{\theta}(t) = (1 - e^{-\frac{2\Phi^\top \Phi}{n} t})\Phi^+ Y$. Since K is invertible with high probability (due to Lemma I.11), the quantity of interest can be expressed as

$$y_i - \varphi(x_i)^\top \hat{\theta}(t) = \varphi(x_i)^\top e^{-\frac{2\Phi^\top \Phi}{n} t} \Phi^+ Y.$$

Our approach consists in iteratively controlling this quantity introducing auxiliary *leave-one-out* variables, *i.e.*, $\Phi_{-i} \in \mathbb{R}^{(n-1)\times p}$ and $Y_{-i} \in \mathbb{R}^{n-1}$, defined without the i-th sample. The first step involves showing that

$$\left\|\Phi^+ Y - \Phi_{-i}^+ Y_{-i}\right\|_2 = \tilde{O}\left(p^{-1/2}\right),$$

which implies that we can focus the analysis on $\sup_{t\in[0,\tau]} |\varphi(x_i)^\top e^{-\frac{2\Phi^\top\Phi}{n}t}\Phi_{-i}^+ Y_{-i}|$. This result comes from the *stability* of GD, as shown in Lemma 8.14, and is a consequence of lower bounding the smallest eigenvalue of the kernel $\lambda_{\min}(K)$.

The second step aims to introduce a leave-one-out variable in the exponent, as we show that

$$\sup_{t\in[0,\tau]} \left|\varphi(x_i)^\top e^{-\frac{2\Phi^\top\Phi}{n}t}\Phi_{-i}^+ Y_{-i}\right| \le 2 \sup_{t\in[0,\tau]} \left|\varphi(x_i)^\top e^{-\frac{2\Phi_{-i}^\top\Phi_{-i}}{n}t}\Phi_{-i}^+ Y_{-i}\right|.$$

This result is achieved via an explicit computation based on *Lie's product formula* for the matrix exponential, which is carried through in Lemma 8.16.

Finally, since the only dependence on x_i is via the term $\varphi(x_i)^\top$, we can conclude the argument with *Dudley's (chaining tail) inequality*, upper bounding the covering number of the set described by the curve $\gamma(t) = V^\top e^{-\frac{2\Phi_{-i}^\top\Phi_{-i}}{n}t}\Phi_{-i}^+ Y_{-i}$ in Lemma 8.17.

Next, we handle the term $|\varphi(x_i)^\top\tilde{\Theta}(t)|$ via the lemma below, also proven in Section 8.6.2.

Lemma 8.11. *Let Assumptions 8.5, 8.6, and 8.8 hold. Then, we have that, jointly for all* $i \in [n]$*,*

$$\sup_{t\in[0,\tau]} \left|\varphi(x_i)^\top\tilde{\Theta}(t)\right| = O(\log n), \tag{8.22}$$

with probability at least $1 - 2\exp\left(-c\log^2 n\right) - 2n\exp(-cp)$ *over* B *and* V*, where* c *is an absolute constant and* B *refers to the probability space of the private mechanism,* i.e.*, the noise in* (8.15)*.*

To prove (8.22), we start by noticing that $\varphi(x_i)^\top\tilde{\Theta}(t)$ evolves as a Gaussian random variable with time-dependent variance. The idea is to upper bound this variance with that of the auxiliary process

$$\mathrm{d}z_i(t) = \varphi(x_i)^\top\Sigma\,\mathrm{d}B(t),$$

which is obtained by removing the *attractive drift* $[-2\Phi^\top(\Phi\hat{\Theta}(t) - Y)/n]$ from (8.15). Intuitively, the removal of the attractive drift increases the variance, as the process becomes less concentrated around the mean. This is formalized in Lemma 8.18, where an application of the *Sudakov-Fernique inequality* gives that

$$\mathbb{E}_B\left[\sup_{t\in[0,\tau]}\left|\varphi(x_i)^\top\tilde{\Theta}(t)\right|\right] \le \mathbb{E}_{z_i}\left[\sup_{t\in[0,\tau]}|z_i(t)|\right]. \tag{8.23}$$

We note that the RHS of (8.23) is easier to control than the LHS, as $z_i(t)$ is a Wiener process and, therefore, its variance is bounded by $O\left(\Sigma^2\tau\,\|\varphi(x_i)\|_2^2\right)$ via the *reflection principle*. Then, given our hyper-parameter choice in (8.13) and the lower bound on ε in Assumption 8.8, we have that $\Sigma^2\tau\,\|\varphi(x_i)\|_2^2 = O(1)$. To conclude, we exploit the *Borell-TIS inequality* to show that, with high probability,

$$\sup_{t\in[0,\tau]}|\varphi(x_i)^\top\tilde{\Theta}(t)| \le \mathbb{E}_B[\sup_{t\in[0,\tau]}|\varphi(x_i)^\top\tilde{\Theta}(t)|] + \log n,$$

which concludes the argument.

The combination of Lemma 8.10 and 8.11 gives that the event in (8.17) happens with high probability, which allows us to study the utility of $\Theta(\tau)$ through the closed form of the OU process $\hat{\Theta}(\tau)$, as $\Theta(\tau) = \hat{\Theta}(\tau)$. This, in turn, boils down to controlling the effects of the noise and of the early stopping that are decoupled via the decomposition $\hat{\Theta}(\tau) = \hat{\theta}(\tau) + \tilde{\Theta}(\tau)$. In fact, on the one hand, $\tilde{\Theta}(\tau)$ is a mean-0 random variable in the probability space of B and it captures the effect of the noise; on the other hand, $\hat{\theta}(\tau)$ is the deterministic component (with respect to B) describing the flow and it captures the effect of the early stopping. We show that both the noise and the early stopping provide negligible damage to utility in the following two lemmas, whose proofs are contained in Section 8.6.3.

Lemma 8.12. *Let Assumptions 8.5, 8.6 and 8.8 hold, and let $d = o(p)$. Then, we have*

$$\mathbb{E}_{x\sim\mathcal{P}_X}\left[\left(\varphi(x)^\top\tilde{\Theta}(\tau)\right)^2\right] = O\left(\frac{d^2\log^{10} n}{n^2}\,\frac{\log(1/\delta)}{\varepsilon^2}\right) = \tilde{O}\left(\frac{d^2}{\varepsilon^2 n^2}\right) = o(1), \tag{8.24}$$

with probability at least $1 - 2\exp\left(-c\log^2 n\right)$ over V and B, where c is an absolute constant.

Similarly to Lemma 8.11, Lemma 8.12 also uses that $\varphi(x_i)^\top\tilde{\Theta}(\tau)$ is a Gaussian random variable with variance increasing linearly in $\|\varphi(x_i)\|_2^2$, τ, and Σ^2. Then, the claim in (8.24) is a consequence of the choice of the hyper-parameters in (8.13) and Assumption 8.8.

Lemma 8.13. *Let Assumptions 8.5, 8.6 and 8.7 hold. Then, we have*

$$\mathbb{E}_{x\sim\mathcal{P}_X}\left[\left(\varphi(x)^\top\left(\hat{\theta}(\tau) - \theta^*\right)\right)^2\right] = O\left(\frac{d}{n} + \frac{n\log^3 d}{d^{3/2}}\right) = \tilde{O}\left(\frac{d}{n} + \frac{n}{d^{3/2}}\right) = o(1), \tag{8.25}$$

with probability at least $1 - 2\exp\left(-c\log^2 n\right)$ over X and V, where c is an absolute constant.

The idea of the argument is to decompose the LHS of (8.25) through two disjoint subspaces, *i.e.*, $\varphi(x)^\top(P_\Lambda + P_\Lambda^\perp)(\hat{\theta}(\tau) - \theta^*)$, where $P_\Lambda \in \mathbb{R}^{p\times p}$ is the projector on the space spanned by the eigenvectors associated with the d largest eigenvalues of $\Phi^\top\Phi$. The rationale is that there is a *spectral gap* between the d-th and the $(d+1)$-th eigenvalue of the kernel K, when sorting them in non-increasing order, see Lemma I.11. We note that this result also uses the well conditioning of the data covariance ($\lambda_{\min}\left(\mathbb{E}_{x\sim\mathcal{P}_X}\left[xx^\top\right]\right) = \Omega(1)$), see the discussion right after the proof of Lemma 8.13 in Section 8.6. As a consequence of the spectral gap, the term $\|P_\Lambda(\hat{\theta}(\tau) - \theta^*)\|_2$ is negligible, since in this subspace $\hat{\theta}(\tau)$ is *already close to convergence*, despite the early stopping. To control the other subspace, we consider the decomposition $\phi(Vx) = \mu_1 Vx + \tilde{\phi}(Vx)$, where μ_1 is the first Hermite coefficient of ϕ. Then, we show that *(i)* $\mathbb{E}_x[(\tilde{\phi}(Vx)^\top P_\Lambda^\perp(\hat{\theta}(\tau) - \theta^*))^2] = O(d/n + n\log^3 d/d^{3/2})$, exploiting the upper bound on $\|\mathbb{E}_x[\tilde{\phi}(Vx)\tilde{\phi}(Vx)^\top]\|_{\mathsf{op}}$ in Lemma I.21, and *(ii)* $\mathbb{E}_x[(x^\top V^\top P_\Lambda^\perp(\hat{\theta}(\tau) - \theta^*))^2] = O\left(d/n + n\log^3 d/d^{3/2}\right)$ exploiting the bound on $\|V^\top P_\Lambda^\perp\Phi^+\|_{\mathsf{op}}$ in Lemma I.20.

Finally, denoting by $\hat{\mathcal{R}}$ and $\mathcal{R}^*$ the generalization error of $\hat{\Theta}(\tau)$ and θ^* respectively, Lemmas 8.12-8.13 (together with some additional manipulations) guarantee that

$$|\hat{\mathcal{R}} - \mathcal{R}^*| = \tilde{O}\left(\frac{d}{n\varepsilon} + \sqrt{\frac{d}{n}} + \sqrt{\frac{n}{d^{3/2}}}\right).$$

As $\Theta(\tau) = \hat{\Theta}(\tau)$ (due to Lemmas 8.10-8.11), the result of Theorem 8.9 follows. The details are deferred to the end of Section 8.6.3.

8.6.2 Analysis of clipping

This section contains the proofs of Lemmas 8.10 and 8.11, including a number of auxiliary results required by the argument. We make use of the notation introduced in the previous section, including the leave-one-out variables ($\Phi_{-1} \in \mathbb{R}^{(n-1)\times p}$, $Y_{-1} \in \mathbb{R}^{n-1}$, $X_{-1} \in \mathbb{R}^{(n-1)\times d}$, $K_{-1} \in \mathbb{R}^{(n-1)\times(n-1)}$) and the OU process $\hat{\Theta}(\tau) = \hat{\theta}(\tau) + \tilde{\Theta}(\tau)$. We also consider the hyper-parameter choice in (8.13). When stating that a random variable Z is sub-Gaussian (sub-exponential), we implicitly mean $\|Z\|_{\psi_2} = O(1)$ ($\|Z\|_{\psi_1} = O(1)$), *i.e.*, its sub-Gaussian (sub-exponential) norm does not increase with the scalings of the problem. Given a p.s.d. matrix A, $\lambda_j(A)$ denotes its j-th eigenvalue sorted in non-increasing order ($\lambda_{\max}(A) = \lambda_1(A) \geq \lambda_2(A) \geq \ldots \geq \lambda_s(A) = \lambda_{\min}(A)$). We indicate with $P_\Lambda \in \mathbb{R}^{p\times p}$ the projector on the space spanned by the eigenvectors associated with the d largest eigenvalues of $\Phi^\top\Phi$. This quantity is well defined due to Lemma I.11 which proves a spectral gap, and we will condition on the result of this lemma throughout the chapter. It is convenient to define the function $\tilde{\phi}(z) := \phi(z) - \mu_1 z$, where μ_1 is the first Hermite coefficient of ϕ. Note that this function is Lipschitz continuous and has the first 3 Hermite coefficients equal to 0, due to Assumption 8.6. We then define the shorthands $\tilde{\varphi}(x) = \tilde{\phi}(Vx) \in \mathbb{R}^p$, $\tilde{\Phi} \in \mathbb{R}^{n\times p}$ as the matrix containing $\tilde{\varphi}(x_i) \in \mathbb{R}^p$ in its i-th row, and $\tilde{K} = \tilde{\Phi}\tilde{\Phi}^\top \in \mathbb{R}^{n\times n}$.

Lemma 8.14. *Let Assumptions 8.5 and 8.6 hold, and let $n = O\left(\sqrt{p}\right)$, $n\log^3 n = O\left(d^{3/2}\right)$ and $n = \omega(d)$. Then, we have*

$$\left\|\Phi^+ Y - \Phi_{-1}^+ Y_{-1}\right\|_2 = O\left(\frac{\log n}{\sqrt{p}}\right), \tag{8.26}$$

with probability at least $1 - 2\exp\left(-c\log^2 n\right)$ over V and X, where c is an absolute constant.

Proof. From Lemma I.11, we have that K is invertible with probability at least $1-2\exp\left(-c_1\log^2 n\right)$ over V and X. Conditioning on such high probability event, we have that also K_{-1} is invertible and, therefore, we can write $\Phi^+ = \Phi^\top K^{-1}$ and $\Phi_{-1}^+ = \Phi_{-1}^\top K_{-1}^{-1}$. Thus, from the proof of Lemma 4.1 in Bombari and Mondelli [2024a] (see their Equation (44)), we have

$$\Phi^+ Y - \Phi_{-1}^+ Y_{-1} = \frac{P_{\Phi_{-1}}^\perp \varphi(x_1)}{\left\|P_{\Phi_{-1}}^\perp \varphi(x_1)\right\|_2^2}\left(y_1 - \varphi(x_1)^\top \Phi_{-1}^+ Y_{-1}\right), \tag{8.27}$$

where $P_{\Phi_{-1}}$ is the projector over the span of the rows of Φ_{-1}.

To bound the term $\varphi(x_1)^\top \Phi_{-1}^+ Y_{-1}$, we decompose $\varphi(x_1) = \mu_1 V x_1 + \tilde{\varphi}(x_1)$. Thus, an application of the triangle inequality gives

$$\left|\varphi(x_1)^\top \Phi_{-1}^+ Y_{-1}\right| \leq \left|\mu_1 x_1^\top V^\top \Phi_{-1}^+ Y_{-1}\right| + \left|\tilde{\varphi}(x_1)^\top \Phi_{-1}^+ Y_{-1}\right|. \tag{8.28}$$

We bound the two terms in (8.28) separately. As for the first term, we have that

$$\left\|V^\top \Phi_{-1}^+ Y_{-1}\right\|_2 \leq \left\|V^\top \Phi_{-1}^+\right\|_{\mathsf{op}} \|Y_{-1}\|_2 = O(1),$$

where the second step is a consequence of Lemma I.16, and holds with probability at least $1 - 2\exp\left(-c_2\log^2 n\right)$ over V and X_{-1}. In fact, notice that considering Φ or Φ_{-1} in Lemma I.16 does not change the argument, and therefore the final result. Then, conditioning on this high probability event, since x_1 is sub-Gaussian, we have that the first term of (8.28) reads

$$\left|\mu_1 x_1^\top V^\top \Phi_{-1}^+ Y_{-1}\right| = O(\log n), \tag{8.29}$$

with probability at least $1 - 2\exp\left(-c_3 \log^2 n\right)$ over V and X.

As for the second term in (8.28), by triangle inequality, we have that

$$\left\|\left(\Phi_{-1}^{+}\right)^{\top} \tilde{\varphi}(x_1)\right\|_2 \leq \left\|K_{-1}^{-1}\mathbb{E}_V\left[\Phi_{-1}\tilde{\varphi}(x_1)\right]\right\|_2 + \left\|K_{-1}^{-1}\left(\Phi_{-1}\tilde{\varphi}(x_1) - \mathbb{E}_V\left[\Phi_{-1}\tilde{\varphi}(x_1)\right]\right)\right\|_2. \quad (8.30)$$

We bound the two terms in (8.30) separately. As for the first term, using indices $i \in [n-1]$, we have

$$\mathbb{E}_V\left[\Phi_{-1}\tilde{\varphi}(x_1)\right]_i = p\,\mathbb{E}_v\left[\phi(x_{i+1}^{\top}v)\tilde{\phi}(x_1^{\top}v)\right] = p\sum_{l=3}^{+\infty}\mu_l^2\left(\frac{x_{i+1}^{\top}x_1}{d}\right)^l, \quad (8.31)$$

where we use the Hermite decomposition of the functions ϕ and $\tilde{\phi}$. Recall that the first 3 Hermite coefficients of $\tilde{\phi}$ are 0, and the others correspond to the ones of ϕ, which we denote as μ_l. Since x_1 is sub-Gaussian and independent from the x_{i+1}-s, with $\|x_{i+1}\|_2 = \sqrt{d}$, we have that

$$\max_{i\in[n-1]}\left|x_{i+1}^{\top}x_1\right| \leq \sqrt{d}\log n, \quad (8.32)$$

with probability at least $1 - 2\exp\left(-c_4 \log^2 n\right)$ over x_1. Thus, conditioning on this high probability event, (8.31) gives

$$\left|\mathbb{E}_V\left[\Phi_{-1}\tilde{\varphi}(x_1)\right]_i\right| \leq p\left(\frac{\left|x_{i+1}^{\top}x_1\right|}{d}\right)^3\sum_{l=3}^{+\infty}\mu_l^2 = O\left(p\frac{\log^3 n}{d^{3/2}}\right), \quad (8.33)$$

which implies

$$\left\|K_{-1}^{-1}\mathbb{E}_V\left[\Phi_{-1}\tilde{\varphi}(x_1)\right]\right\|_2 \leq \left\|K_{-1}^{-1}\right\|_{\mathsf{op}}\left\|\mathbb{E}_V\left[\Phi_{-1}\tilde{\varphi}(x_1)\right]\right\|_2 = O\left(\frac{\sqrt{n}\log^3 n}{d^{3/2}}\right), \quad (8.34)$$

with probability at least $1 - 2\exp\left(-c_5 \log^2 n\right)$ over X and V. Here, the last passage also uses Lemma I.11, which provides a lower bound on $\lambda_{\min}\left(K_{-1}\right) \geq \lambda_{\min}\left(K\right) = \Omega(p)$.

As for the second term of (8.30), we have

$$\left[\Phi_{-1}\tilde{\varphi}(x_1)\right]_i - \mathbb{E}_V\left[\Phi_{-1}\tilde{\varphi}(x_1)\right]_i = \sum_{k=1}^{p}\phi(x_{i+1}^{\top}v_k)\tilde{\phi}(x_1^{\top}v_k) - \mathbb{E}_{v_k}\left[\phi(x_{i+1}^{\top}v_k)\tilde{\phi}(x_1^{\top}v_k)\right]. \quad (8.35)$$

All terms in the previous sum are independent, sub-exponential, mean-0 random variables, since both ϕ and $\tilde{\phi}$ are Lipschitz. Thus, Bernstein inequality (see Theorem 2.8.1 of Vershynin [2018]) gives

$$\left|\left[\Phi_{-1}\tilde{\varphi}(x_1)\right]_i - \mathbb{E}_V\left[\Phi_{-1}\tilde{\varphi}(x_1)\right]_i\right| = O\left(\log n\sqrt{p}\right), \quad (8.36)$$

with probability at least $1 - 2\exp\left(-c_6 \log^2 n\right)$ over V. Then, performing a union bound over all the indices $i \in [n-1]$, we obtain

$$\left\|K_{-1}^{-1}\left(\Phi_{-1}\tilde{\varphi}(x_1) - \mathbb{E}_V\left[\Phi_{-1}\tilde{\varphi}(x_1)\right]\right)\right\|_2 = O\left(\frac{1}{p}\sqrt{np}\log n\right) = O\left(\sqrt{\frac{n}{p}}\log n\right), \quad (8.37)$$

with probability at least $1 - 2\exp\left(-c_7 \log^2 n\right)$ over V and X. Then, (8.37) and (8.34) make (8.30) read

$$\left\|\left(\Phi_{-1}^{+}\right)^{\top}\tilde{\varphi}(x_1)\right\|_2 = O\left(\frac{\sqrt{n}\log^3 n}{d^{3/2}} + \sqrt{\frac{n}{p}}\log n\right) = O\left(\frac{\log n}{\sqrt{n}}\right), \quad (8.38)$$

with probability at least $1 - 2\exp\left(-c_8 \log^2 n\right)$ over V and X, where the last step is a consequence of $p = \Omega(n^2)$. Plugging (8.38) and (8.29) in (8.28) provides the upper bound

$$\left|\varphi(x_1)^\top \Phi_{-1}^+ Y_{-1}\right| = O\left(\log n\right), \tag{8.39}$$

which holds with probability at least $1 - 2\exp\left(-c_9 \log^2 n\right)$ over V and X.

To bound the term $\left\|P_{\Phi_{-1}}^\perp \varphi(x_1)\right\|_2$ in (8.27), we can use Lemma B.1. in Bombari and Mondelli [2024a], which gives

$$\left\|P_{\Phi_{-1}}^\perp \varphi(x_1)\right\|_2 \geq \sqrt{\lambda_{\min}\left(K\right)} = \Omega(\sqrt{p}), \tag{8.40}$$

where the last step is a consequence of Lemma I.11, and holds with probability at least $1 - 2\exp\left(-c_{10} \log^2 n\right)$ over V and X. The combination of (8.27), (8.39) and (8.40) readily gives the desired result. □

After performing a union bound on all $i \in [n]$, the result of Lemma 8.14 guarantees that, with high probability,

$$\left\|\Phi^+ Y - \Phi_{-i}^+ Y_{-i}\right\|_2 = O\left(\frac{\log n}{\sqrt{p}}\right), \qquad \text{for all } i \in [n].$$

This will be key in our later proof of Lemma 8.10, where we will use Lemma 8.14 and the fact that it holds with probability at least $1 - 2\exp\left(-c \log^2 n\right)$. We note that the proof strategy of Lemma 8.14 is specifically designed to obtain such a probability guarantee, and it differs from previous work, especially in the argument to show

$$\|(\Phi_{-i}^+)^\top \tilde{\varphi}(x_i)\|_2 = O\left(\frac{\log n}{\sqrt{n}}\right), \qquad \text{for all } i \in [n],$$

see (8.38). Note that, for a fixed i, this can be proved via Markov inequality through an upper bound on the second moment of the LHS of the equation above, which in turn reduces to an upper bound on $\|\mathbb{E}_x[\tilde{\varphi}(x)\tilde{\varphi}(x)^\top]\|_{\mathsf{op}}$. This resembles the strategy followed by Mei et al. [2021] (see their Proposition 7.(b)) to prove that this term is negligible when estimating the test loss of the RF model, which is their object of interest. However, this approach would lead to vacuous probability guarantees after performing a union bound on all the training samples, due to the polynomial tail bound given by Markov inequality. To solve the issue, we pursue our desired result via the argument between (8.30) and (8.34), which leads to

$$\frac{\sqrt{n}\log^3 n}{d^{3/2}} = O\left(\frac{\log n}{\sqrt{n}}\right),$$

and therefore to our assumption $d^{3/2} \gg n$, see (8.11). Similarly, in the argument between (8.35) and (8.37), we upper bound $\|(\Phi_{-i}^+)^\top \tilde{\varphi}(x_i) - \mathbb{E}_V\left[(\Phi_{-i}^+)^\top \tilde{\varphi}(x_i)\right]\|_2$ via Bernstein inequality on the separate indices of the argument vector, which at the end relies on the bound

$$\sqrt{\frac{n}{p}\log n} = O\left(\frac{\log n}{\sqrt{n}}\right),$$

leading to our assumption $p = \Omega(n^2)$, see (8.11).

Lemma 8.15. *Let Assumptions 8.5 and 8.6 hold, and let* $n = o\left(p/\log^4 p\right)$, $n = \omega\left(d\log^2 d\right)$, *and* $n = O\left(d^{3/2}/\log^3 d\right)$. *Then, we have that, with probability at least* $1 - 2\exp\left(-c\log^2 n\right)$ *over* X *and* V,

$$\frac{\varphi(x_1)^\top}{\|\varphi(x_1)\|_2} e^{-\frac{2\Phi^\top\Phi}{n}t} \frac{\varphi(x_1)}{\|\varphi(x_1)\|_2} < \frac{\varphi(x_1)^\top}{\|\varphi(x_1)\|_2} e^{-\frac{2\varphi(x_1)\varphi(x_1)^\top}{n}t} \frac{\varphi(x_1)}{\|\varphi(x_1)\|_2}, \tag{8.41}$$

holds uniformly for all $t \in (0, \tau]$, *where* c *is an absolute constant.*

Proof. Let P_Λ be the projector on the space spanned by the eigenvectors associated with the d largest eigenvalues of $\Phi^\top\Phi$, P_0 be the projector on the kernel of $\Phi^\top\Phi$, and $P_\lambda = I - P_\Lambda - P_0$. By definition, we have that $P_0\varphi(x_1) = 0$. Thus, since both P_Λ and P_λ are projectors on eigenspaces of $\Phi^\top\Phi$, we can write

$$\begin{aligned}
&\frac{\varphi(x_1)^\top}{\|\varphi(x_1)\|_2} e^{-\frac{2\Phi^\top\Phi}{n}t} \frac{\varphi(x_1)}{\|\varphi(x_1)\|_2} \\
&= \frac{\varphi(x_1)^\top P_\lambda}{\|\varphi(x_1)\|_2} e^{-\frac{2P_\lambda\Phi^\top\Phi P_\lambda}{n}t} \frac{P_\lambda\varphi(x_1)}{\|\varphi(x_1)\|_2} + \frac{\varphi(x_1)^\top P_\Lambda}{\|\varphi(x_1)\|_2} e^{-\frac{2P_\Lambda\Phi^\top\Phi P_\Lambda}{n}t} \frac{P_\Lambda\varphi(x_1)}{\|\varphi(x_1)\|_2} \\
&\le e^{-\lambda t} \frac{\|P_\lambda\varphi(x_1)\|_2^2}{\|\varphi(x_1)\|_2^2} + e^{-\Lambda t} \frac{\|P_\Lambda\varphi(x_1)\|_2^2}{\|\varphi(x_1)\|_2^2},
\end{aligned} \tag{8.42}$$

where we have defined the shorthands

$$\Lambda = \frac{2\lambda_d(K)}{n}, \qquad \lambda = \frac{2\lambda_{\min}(K)}{n}. \tag{8.43}$$

Note that the last step of (8.42) holds as $e^{-\lambda t}$ and $e^{-\Lambda t}$ are the largest eigenvalues of $e^{-\frac{2\Phi^\top\Phi}{n}t}$ in the subspaces P_λ and P_Λ are respectively projecting on. Furthermore, the RHS of (8.41) reads

$$\frac{\varphi(x_1)^\top}{\|\varphi(x_1)\|_2} e^{-\frac{2\varphi(x_1)\varphi(x_1)^\top}{n}t} \frac{\varphi(x_1)}{\|\varphi(x_1)\|_2} = e^{-\frac{2\|\varphi(x_1)\|_2^2}{n}t}. \tag{8.44}$$

In remaining part of the proof, we use the following inequality

$$1 - \frac{cz}{e} \ge e^{-cz}, \tag{8.45}$$

which holds for any $c > 0$, and $0 < z \le 1/c$. We also use that, for all $z > 0$,

$$1 - cz < e^{-cz}. \tag{8.46}$$

We prove the inequality in (8.41) in two disjoint intervals.

1. $0 < t \le 1/\Lambda$: In this interval, we can use (8.42) and apply (8.45) twice (note that $1/\Lambda \le 1/\lambda$), obtaining

$$\begin{aligned}
\frac{\varphi(x_1)^\top}{\|\varphi(x_1)\|_2} e^{-\frac{2\Phi^\top\Phi}{n}t} \frac{\varphi(x_1)}{\|\varphi(x_1)\|_2} &\le e^{-\lambda t} \frac{\|P_\lambda\varphi(x_1)\|_2^2}{\|\varphi(x_1)\|_2^2} + e^{-\Lambda t} \frac{\|P_\Lambda\varphi(x_1)\|_2^2}{\|\varphi(x_1)\|_2^2} \\
&\le \left(1 - \frac{\Lambda t}{e}\right) \frac{\|P_\Lambda\varphi(x_1)\|_2^2}{\|\varphi(x_1)\|_2^2} + \left(1 - \frac{\lambda t}{e}\right) \frac{\|P_\lambda\varphi(x_1)\|_2^2}{\|\varphi(x_1)\|_2^2} \\
&= 1 - \frac{\Lambda t}{e} \frac{\|P_\Lambda\varphi(x_1)\|_2^2}{\|\varphi(x_1)\|_2^2} - \frac{\lambda t}{e} \frac{\|P_\lambda\varphi(x_1)\|_2^2}{\|\varphi(x_1)\|_2^2}.
\end{aligned} \tag{8.47}$$

Applying (8.46) to (8.44) we obtain

$$\frac{\varphi(x_1)^\top}{\|\varphi(x_1)\|_2} e^{-\frac{2\varphi(x_1)\varphi(x_1)^\top}{n}t} \frac{\varphi(x_1)}{\|\varphi(x_1)\|_2} > 1 - \frac{2\|\varphi(x_1)\|_2^2}{n}t. \tag{8.48}$$

Then, on this interval, (8.47) and (8.48) imply that proving the following

$$1 - \frac{\Lambda t}{e}\frac{\|P_\Lambda \varphi(x_1)\|_2^2}{\|\varphi(x_1)\|_2^2} - \frac{\lambda t}{e}\frac{\|P_\lambda \varphi(x_1)\|_2^2}{\|\varphi(x_1)\|_2^2} \overset{?}{\leq} 1 - \frac{2\|\varphi(x_1)\|_2^2}{n}t, \tag{8.49}$$

is enough to prove the thesis. This, in turn, can be shown proving that

$$\Lambda = \frac{2\lambda_d(K)}{n} \overset{?}{\geq} \frac{2e\|\varphi(x_1)\|_2^4}{n\|P_\Lambda \varphi(x_1)\|_2^2}, \tag{8.50}$$

where the ? remarks that this inequality still has to be proved.

Now, by Lemmas I.11 and I.18, we jointly have

$$\lambda_d(K) = \Omega\left(\frac{pn}{d}\right) = \Omega(p\log n), \qquad \|P_\Lambda \varphi(x_1)\|_2^2 = \Omega(p), \tag{8.51}$$

with probability at least $1 - 2\exp\left(-c_2 \log^2 n\right)$ over X and V. Then, conditioning on this high probability event, the LHS of (8.50) is $\Omega(p\log n\,/n)$, while its RHS is $O\left(p/n\right)$ (recall that $\|\varphi(x_1)\|_2^2 = O\left(p\right)$). Thus, (8.50) holds for n sufficiently large, which gives that the desired result holds in the interval $0 < t \leq 1/\Lambda$, with probability at least $1 - 2\exp\left(-c_3 \log^2 n\right)$ over X and V (where c_3 is eventually smaller than c_2, to make this probability being 0 for the n-s that are not large enough).

2. $\underline{1/\Lambda < t \leq \tau}$: We have

$$\begin{aligned}
\frac{\varphi(x_1)^\top}{\|\varphi(x_1)\|_2} e^{-\frac{2\Phi^\top\Phi}{n}t} \frac{\varphi(x_1)}{\|\varphi(x_1)\|_2} &\leq e^{-\lambda t}\frac{\|P_\lambda \varphi(x_1)\|_2^2}{\|\varphi(x_1)\|_2^2} + e^{-\Lambda t}\frac{\|P_\Lambda \varphi(x_1)\|_2^2}{\|\varphi(x_1)\|_2^2} \\
&\leq \frac{\|P_\lambda \varphi(x_1)\|_2^2}{\|\varphi(x_1)\|_2^2} + e^{-1}\frac{\|P_\Lambda \varphi(x_1)\|_2^2}{\|\varphi(x_1)\|_2^2} \\
&= 1 - \left(1 - e^{-1}\right)\frac{\|P_\Lambda \varphi(x_1)\|_2^2}{\|\varphi(x_1)\|_2^2}.
\end{aligned} \tag{8.52}$$

Applying (8.46) to (8.44) we obtain

$$\frac{\varphi(x_1)^\top}{\|\varphi(x_1)\|_2} e^{-\frac{2\varphi(x_1)\varphi(x_1)^\top}{n}t} \frac{\varphi(x_1)}{\|\varphi(x_1)\|_2} > 1 - \frac{2\|\varphi(x_1)\|_2^2}{n}t \geq 1 - \frac{2\|\varphi(x_1)\|_2^2}{n}\tau. \tag{8.53}$$

Then, on this interval, (8.52) and (8.53) imply that proving the following

$$\left(1 - e^{-1}\right)\frac{\|P_\Lambda \varphi(x_1)\|_2^2}{\|\varphi(x_1)\|_2^2} \overset{?}{\geq} \frac{2\|\varphi(x_1)\|_2^2}{n}\tau, \tag{8.54}$$

is enough to prove the thesis. By Lemma I.18, we have that the LHS of the previous equation is $\Theta(1)$ with probability at least $1 - 2\exp\left(-c_4 \log^2 n\right)$ over X and V. For

the RHS, since $\|\varphi(x_1)\|_2^2 = O(p)$ with probability at least $1 - 2\exp(-c_5 p)$ over V (see the argument carried out in (I.147) and (I.148)), we have that

$$\frac{2\|\varphi(x_1)\|_2^2}{n}\tau = O\left(\frac{p}{n}\frac{d\log^2 n}{p}\right) = o(1). \tag{8.55}$$

Thus, (8.54) holds for n sufficiently large, which gives that the desired result holds in the interval $1/\Lambda < t \le \tau$, with probability at least $1 - 2\exp\left(-c_6 \log^2 n\right)$ over X and V (where c_6 is set to make this probability being 0 for the n-s that are not large enough).

□

Lemma 8.16. *Let Assumptions 8.5 and 8.6 hold, and let* $n = o\left(p/\log^4 p\right)$, $n = \omega\left(d\log^2 d\right)$, *and* $n = O\left(d^{3/2}/\log^3 d\right)$. *Then, we have*

$$\sup_{t\in[0,\tau]}\left|\varphi(x_1)^\top e^{-\frac{2\Phi^\top\Phi}{n}t}\Phi_{-1}^+ Y_{-1}\right| \le 2\sup_{t\in[0,\tau]}\left|\varphi(x_1)^\top e^{-\frac{2\Phi_{-1}^\top\Phi_{-1}}{n}t}\Phi_{-1}^+ Y_{-1}\right|, \tag{8.56}$$

with probability at least $1 - 2\exp\left(-c\log^2 n\right)$ *over* X *and* V, *where* c *is an absolute constant.*

Proof. Note that, for any positive number s,

$$e^{-\frac{2\varphi(x_1)\varphi(x_1)^\top}{ns}t} = e^{-\frac{\|2\varphi(x_1)\|_2^2}{ns}t}\frac{\varphi(x_1)\varphi(x_1)^\top}{\|\varphi(x_1)\|_2^2} + P_{\varphi(x_1)}^\perp, \tag{8.57}$$

where $P_{\varphi(x_1)}^\perp \in \mathbb{R}^{p\times p}$ is the orthogonal projector to the space spanned by $\varphi(x_1)$, *i.e.*, $P_{\varphi(x_1)}^\perp = I - \varphi(x_1)\varphi(x_1)^\top/\|\varphi(x_1)\|_2^2$. Thus, we can write

$$e^{-\frac{2\varphi(x_1)\varphi(x_1)^\top}{ns}t} = I + \alpha(s)\frac{\varphi(x_1)\varphi(x_1)^\top}{\|\varphi(x_1)\|_2^2}, \tag{8.58}$$

where we introduced the shorthand

$$-1 < \alpha(s) = -\left(1 - e^{-\frac{2\|\varphi(x_1)\|_2^2}{ns}t}\right) < 0. \tag{8.59}$$

Let us also introduce

$$\Pi(s) = \left(e^{-\frac{2\varphi(x_1)\varphi(x_1)^\top}{ns}t}e^{-\frac{2\Phi_{-1}^\top\Phi_{-1}}{ns}t}\right)^s \in \mathbb{R}^{p\times p}, \qquad \chi(s) = \left|\varphi(x_1)^\top\Pi(s)\Phi_{-1}^+ Y_{-1}\right| \in \mathbb{R}, \tag{8.60}$$

defined for any s being a positive natural number. As $\varphi(x_1)\varphi(x_1)^\top + \Phi_{-1}^\top\Phi_{-1} = \Phi^\top\Phi$, an application of Lie's product formula gives

$$\lim_{s\to\infty}\Pi(s) = e^{-\frac{2\Phi^\top\Phi}{n}t}, \tag{8.61}$$

and therefore

$$\lim_{s\to\infty}\chi(s) = \left|\varphi(x_1)^\top e^{-\frac{2\Phi^\top\Phi}{n}t}\Phi_{-1}^+ Y_{-1}\right|. \tag{8.62}$$

Plugging (8.58) in the expression of $\Pi(s)$ gives

$$\Pi(s) = \left(\left(I + \alpha(s)\frac{\varphi(x_1)\varphi(x_1)^\top}{\|\varphi(x_1)\|_2^2}\right) A(s)\right)^s, \tag{8.63}$$

where we define

$$A(s) = e^{-\frac{2\Phi_{-1}^\top \Phi_{-1}}{ns}t} \in \mathbb{R}^{p\times p}. \tag{8.64}$$

Note that $\Pi(s)$ can be expanded as

$$\Pi(s) = \left(I + \alpha(s)\frac{\varphi(x_1)\varphi(x_1)^\top}{\|\varphi(x_1)\|_2^2}\right)\sum_{l=1}^{s-1}\Pi_l(s)A(s)^{s-l} + \left(I + \alpha(s)\frac{\varphi(x_1)\varphi(x_1)^\top}{\|\varphi(x_1)\|_2^2}\right)A(s)^s, \tag{8.65}$$

where we define

$$\Pi_l(s) = \left(A(s)\left(I + \alpha(s)\frac{\varphi(x_1)\varphi(x_1)^\top}{\|\varphi(x_1)\|_2^2}\right)\right)^{l-1} A(s)\alpha(s)\frac{\varphi(x_1)\varphi(x_1)^\top}{\|\varphi(x_1)\|_2^2}. \tag{8.66}$$

In words, $\Pi_l(s)$ includes all the terms where the last term containing $\alpha(s)$ is taken at the $(l+1)$-th factor of (8.63). This gives

$$\begin{aligned}
\chi(s) &= \left|\varphi(x_1)^\top \Pi(s)\Phi_{-1}^+ Y_{-1}\right| \\
&= \left|\left(1+\alpha(s)\right)\varphi(x_1)^\top \sum_{l=1}^{s-1}\Pi_l(s)A(s)^{s-l}\Phi_{-1}^+ Y_{-1} + \left(1+\alpha(s)\right)\varphi(x_1)^\top A(s)^s\Phi_{-1}^+ Y_{-1}\right| \\
&= \left|\left(1+\alpha(s)\right)\alpha(s)\sum_{l=1}^{s-1}\pi_l(s)\varphi(x_1)^\top A(s)^{s-l}\Phi_{-1}^+ Y_{-1} + \left(1+\alpha(s)\right)\varphi(x_1)^\top A(s)^s\Phi_{-1}^+ Y_{-1}\right| \\
&\le \left|\left(1+\alpha(s)\right)\alpha(s)\sum_{l=1}^{s-1}\pi_l(s)\right| \max_{l\in[s-1]}\left|\varphi(x_1)^\top A(s)^{s-l}\Phi_{-1}^+ Y_{-1}\right| + \\
&\quad + \left|\left(1+\alpha(s)\right)\varphi(x_1)^\top A(s)^s\Phi_{-1}^+ Y_{-1}\right|,
\end{aligned} \tag{8.67}$$

where we introduce the shorthand

$$\begin{aligned}
\pi_l(s) &= \frac{\varphi(x_1)^\top}{\|\varphi(x_1)\|_2}\left(A(s)\left(I + \alpha(s)\frac{\varphi(x_1)\varphi(x_1)^\top}{\|\varphi(x_1)\|_2^2}\right)\right)^{l-1} A(s)\frac{\varphi(x_1)}{\|\varphi(x_1)\|_2} \\
&= \frac{\left(A(s)^{1/2}\varphi(x_1)\right)^\top}{\|\varphi(x_1)\|_2} M_l(s)\frac{A(s)^{1/2}\varphi(x_1)}{\|\varphi(x_1)\|_2},
\end{aligned} \tag{8.68}$$

and $M_l(s)$ is the p.s.d. matrix defined as

$$M_l(s) = \left(A(s)^{1/2}\left(I + \alpha(s)\frac{\varphi(x_1)\varphi(x_1)^\top}{\|\varphi(x_1)\|_2^2}\right)A(s)^{1/2}\right)^{l-1}. \tag{8.69}$$

Then, note that (8.62) and (8.67) give

$$\begin{aligned}
&\left|\varphi(x_1)^\top e^{-\frac{2\Phi^\top\Phi}{n}t}\Phi_{-1}^+Y_{-1}\right| \\
&\leq \limsup_{s\to\infty}\left(\left|(1+\alpha(s))\,\alpha(s)\sum_{l=1}^{s-1}\pi_l(s)\right|\max_{l\in[s-1]}\left|\varphi(x_1)^\top e^{-\frac{2\Phi_{-1}^\top\Phi_{-1}}{ns}(s-l)t}\Phi_{-1}^+Y_{-1}\right|\right) \\
&\quad+\limsup_{s\to\infty}\left(|1+\alpha(s)|\left|\varphi(x_1)^\top e^{-\frac{2\Phi_{-1}^\top\Phi_{-1}}{ns}st}\Phi_{-1}^+Y_{-1}\right|\right) \\
&\leq \limsup_{s\to\infty}\left|(1+\alpha(s))\,\alpha(s)\sum_{l=1}^{s-1}\pi_l(s)\right|\sup_{t'\in(0,t)}\left|\varphi(x_1)^\top e^{-\frac{2\Phi_{-1}^\top\Phi_{-1}}{n}t'}\Phi_{-1}^+Y_{-1}\right| \\
&\quad+\limsup_{s\to\infty}|1+\alpha(s)|\left|\varphi(x_1)^\top e^{-\frac{2\Phi_{-1}^\top\Phi_{-1}}{n}t}\Phi_{-1}^+Y_{-1}\right| \\
&\leq \left(1+\limsup_{s\to\infty}\left|\alpha(s)\sum_{l=1}^{s-1}\pi_l(s)\right|\right)\sup_{t'\in(0,t)}\left|\varphi(x_1)^\top e^{-\frac{2\Phi_{-1}^\top\Phi_{-1}}{n}t'}\Phi_{-1}^+Y_{-1}\right|,
\end{aligned} \tag{8.70}$$

where in the last line we used $\lim_{s\to\infty}\alpha(s)=0$, which follows from (8.59). We will now upper bound the first factor of this last expression, for all $t\in(0,\tau]$, as we will treat the case $t=0$ separately. Note that, following (8.68), we have

$$\begin{aligned}
\pi_l(s) &= \frac{\left(A(s)^{1/2}\varphi(x_1)\right)^\top}{\|\varphi(x_1)\|_2}M_l(s)\frac{A(s)^{1/2}\varphi(x_1)}{\|\varphi(x_1)\|_2} \\
&= \frac{\left\|A(s)^{1/2}\varphi(x_1)\right\|_2^2}{\|\varphi(x_1)\|_2^2}\frac{\left(A(s)^{1/2}\varphi(x_1)\right)^\top}{\|A(s)^{1/2}\varphi(x_1)\|_2}M_s(s)^{(l-1)/(s-1)}\frac{A(s)^{1/2}\varphi(x_1)}{\|A(s)^{1/2}\varphi(x_1)\|_2} \\
&\leq \frac{\left\|A(s)^{1/2}\varphi(x_1)\right\|_2^2}{\|\varphi(x_1)\|_2^2}\left(\frac{\left(A(s)^{1/2}\varphi(x_1)\right)^\top}{\|A(s)^{1/2}\varphi(x_1)\|_2}M_s(s)\frac{A(s)^{1/2}\varphi(x_1)}{\|A(s)^{1/2}\varphi(x_1)\|_2}\right)^{(l-1)/(s-1)} \\
&=: \frac{\left\|A(s)^{1/2}\varphi(x_1)\right\|_2^2}{\|\varphi(x_1)\|_2^2}\mu(s)^{(l-1)/(s-1)},
\end{aligned} \tag{8.71}$$

where the second line follows directly from (8.69), and the third line is a consequence of Jensen inequality, since $M_s(s)$ is p.s.d. and $l\leq s$.

From (8.64), we have that $\lim_{s\to\infty}A(s)=I$. Thus,

$$\begin{aligned}
\lim_{s\to\infty}M_s(s) &= \lim_{s\to\infty}\left(A(s)^{1/2}\left(I+\alpha(s)\frac{\varphi(x_1)\varphi(x_1)^\top}{\|\varphi(x_1)\|_2^2}\right)A(s)^{1/2}\right)^{s-1} \\
&= \lim_{s\to\infty}A(s)^{1/2}\Pi(s)\left(\left(I+\alpha(s)\frac{\varphi(x_1)\varphi(x_1)^\top}{\|\varphi(x_1)\|_2^2}\right)A(s)\right)^{-1}A(s)^{1/2} \\
&= \lim_{s\to\infty}\Pi(s),
\end{aligned} \tag{8.72}$$

and

$$\begin{aligned}\lim_{s\to\infty}\mu(s) &= \lim_{s\to\infty}\frac{\left(A(s)^{1/2}\varphi(x_1)\right)^\top}{\|A(s)^{1/2}\varphi(x_1)\|_2}\lim_{s\to\infty}M_s(s)\lim_{s\to\infty}\frac{A(s)^{1/2}\varphi(x_1)}{\|A(s)^{1/2}\varphi(x_1)\|_2}\\ &= \frac{\varphi(x_1)^\top}{\|\varphi(x_1)\|_2}\left(\lim_{s\to\infty}\Pi(s)\right)\frac{\varphi(x_1)}{\|\varphi(x_1)\|_2}\\ &= \frac{\varphi(x_1)^\top}{\|\varphi(x_1)\|_2}e^{-\frac{2\Phi^\top\Phi}{n}t}\frac{\varphi(x_1)}{\|\varphi(x_1)\|_2}\\ &< \frac{\varphi(x_1)^\top}{\|\varphi(x_1)\|_2}e^{-\frac{2\varphi(x_1)\varphi(x_1)^\top}{n}t}\frac{\varphi(x_1)}{\|\varphi(x_1)\|_2}\\ &= e^{-\frac{2\|\varphi(x_1)\|_2^2}{n}t},\end{aligned} \tag{8.73}$$

where the third line follows from (8.61) and Lemma 8.15 guarantees that, with probability at least $1-2\exp\left(-c\log^2 n\right)$ over X and V, the fourth line uniformly holds for all $t\in(0,\tau)$. We will condition on this event until the end of the proof.

The previous limit implies that there exists s^* such that, for all $s>s^*$, we have $\mu(s)<e^{-\frac{2\|\varphi(x_1)\|_2^2}{n}t}$. Thus, for such s, we also have

$$0<\mu(s)^{1/(s-1)}<\mu(s)^{1/s}\leq e^{-\frac{2\|\varphi(x_1)\|_2^2}{ns}t}=1+\alpha(s)<1. \tag{8.74}$$

Then, for such s, (8.71) leads to

$$\begin{aligned}\sum_{l=1}^{s-1}\pi_l(s) &\leq \frac{\left\|A(s)^{1/2}\varphi(x_1)\right\|_2^2}{\|\varphi(x_1)\|_2^2}\sum_{l=1}^{s-1}\mu(s)^{(l-1)/(s-1)}\\ &= \frac{\left\|A(s)^{1/2}\varphi(x_1)\right\|_2^2}{\|\varphi(x_1)\|_2^2}\frac{1-\mu(s)}{1-\mu(s)^{1/(s-1)}}\\ &< \frac{\left\|A(s)^{1/2}\varphi(x_1)\right\|_2^2}{\|\varphi(x_1)\|_2^2}\frac{1-\mu(s)}{-\alpha(s)},\end{aligned} \tag{8.75}$$

where we solve the geometric series in the second line and use (8.74) in the third one. This gives

$$\begin{aligned}&\limsup_{s\to\infty}\left|\alpha(s)\sum_{l=1}^{s-1}\pi_l(s)\right|\\ &\leq \lim_{s\to\infty}\left|\frac{\alpha(s)}{-\alpha(s)}\right|\lim_{s\to\infty}\frac{\left\|A(s)^{1/2}\varphi(x_1)\right\|_2^2}{\|\varphi(x_1)\|_2^2}\lim_{s\to\infty}|1-\mu(s)|\leq 1.\end{aligned} \tag{8.76}$$

Plugging this last result in (8.70), we get

$$\left|\varphi(x_1)^\top e^{-\frac{2\Phi^\top\Phi}{n}t}\Phi_{-1}^+Y_{-1}\right|\leq 2\sup_{t'\in(0,t)}\left|\varphi(x_1)^\top e^{-\frac{2\Phi_{-1}^\top\Phi_{-1}}{n}t'}\Phi_{-1}^+Y_{-1}\right|, \tag{8.77}$$

which, taking the supremum of $t\in(0,\tau)$, and extending by continuity to $t=0$, leads to the desired result. □

For the next lemma, it is convenient to define the ϵ-covering number of a separable set T as follows

$$\mathcal{N}(T,\epsilon) := \inf\left\{|T_0| \text{ such that } T_0 \subseteq T, \text{ and } T \subseteq \bigcup_{t_0\in T_0} \bar{S}(t_0,\epsilon)\right\}, \tag{8.78}$$

where $|T_0|$ denotes the cardinality of the set T_0, and $\bar{S}(t_0,\epsilon)$ denotes the closed Euclidean ball with center t_0 and radius ϵ. We denote with $\mathrm{diam}(T) = \sup_{u,u'\in T} \|u-u'\|_2$ the diameter of T.

Lemma 8.17. *Let Assumptions 8.5 and 8.6 hold, and let $n = o\left(p/\log^4 p\right)$, $n\log^3 n = O\left(d^{3/2}\right)$ and $n = \omega(d)$. Let $T \subseteq \mathbb{R}^d$ be the set described by the curve $\gamma : \mathbb{R} \to \mathbb{R}^d$ defined as*

$$\gamma(t) = V^\top e^{-\frac{2\Phi_{-1}^\top \Phi_{-1}}{n}t}\Phi_{-1}^+ Y_{-1}, \tag{8.79}$$

for $t \in [0,\tau]$. Then, T is separable with respect to the Euclidean norm and, denoting with $\mathcal{N}(T,\epsilon)$ its ϵ-covering number and with $\mathrm{diam}(T)$ its diameter, we jointly have that

$$\int_0^\infty \sqrt{\log \mathcal{N}(T,\epsilon)}\mathrm{d}\epsilon = O(\log n), \qquad \mathrm{diam}(T) = O(1), \tag{8.80}$$

with probability at least $1-2\exp\left(-c\log^2 n\right)$ over X_{-1} and V, where c is an absolute constant.

Proof. T is described by a continuous function γ applied to $t \in [0,\tau]$. Since the interval $[0,\tau]$ is separable, we also have that T is separable. Furthermore, note that $e^{-\frac{2\Phi_{-1}^\top\Phi_{-1}}{n}t}\Phi_{-1}^+ = \Phi_{-1}^+ e^{-\frac{2K_{-1}}{n}t}$. Then, for all $t,t' \in [0,\tau]$, we have

$$\|\gamma(t)-\gamma(t')\|_2 = \left\|V^\top\Phi_{-1}^+ e^{-\frac{2K_{-1}}{n}t}\left(I - e^{-\frac{2K_{-1}}{n}(t'-t)}\right)Y_{-1}\right\|_2. \tag{8.81}$$

Assuming, without loss of generality, $t' \geq t$, we have that

$$\|\gamma(t)-\gamma(t')\|_2 \leq \left\|V^\top\Phi_{-1}^+\right\|_{\mathsf{op}} \|Y_{-1}\|_2 = O(1), \tag{8.82}$$

where the last step is a consequence of Lemma I.16 and it holds with probability at least $1-2\exp\left(-c_1\log^2 n\right)$ over X_{-1} and V (note that the argument of Lemma I.16 goes through equivalently when considering Φ_{-1} instead of Φ). This proves the desired result on the diameter.

Following a similar strategy, we can prove that, for all $t' > t$ s.t. $|t-t'| \leq \Delta$,

$$\begin{aligned}
\|\gamma(t)-\gamma(t')\|_2 &\leq \left\|V^\top\Phi_{-1}^+\right\|_{\mathsf{op}} \left\|I - e^{-\frac{2K_{-1}}{n}(t'-t)}\right\|_{\mathsf{op}} \|Y_{-1}\|_2 \\
&\leq \left\|V^\top\Phi_{-1}^+\right\|_{\mathsf{op}} \left(1 - e^{-\frac{2\lambda_{\max}(K_{-1})}{n}(t'-t)}\right)\|Y_{-1}\|_2 \\
&\leq \left\|V^\top\Phi_{-1}^+\right\|_{\mathsf{op}} \frac{2\lambda_{\max}(K_{-1})}{n}(t'-t)\|Y_{-1}\|_2 \\
&\leq \left\|V^\top\Phi_{-1}^+\right\|_{\mathsf{op}} \|Y_{-1}\|_2 \frac{2\lambda_{\max}(K_{-1})}{n}\Delta,
\end{aligned} \tag{8.83}$$

where the second step is a consequence of the fact that $1 - e^{-\frac{2K_{-1}}{n}(t'-t)}$ is a p.s.d. matrix with eigenvalues given by $1 - e^{-\frac{2\lambda_i(K_{-1})}{n}(t'-t)}$, and the third step follows from the inequality

$z \geq 1 - e^{-z}$, for $z \geq 0$. Since $\|\varphi(x_i)\|_2^2 = O(p)$ jointly for all $i \in [n]$ with probability at least $1 - 2\exp(-c_2 p)$ over V (see the argument carried out in (I.147) and (I.148)), we have that $\lambda_{\max}(K_{-1}) = \|\Phi_{-1}\|_{\text{op}}^2 \leq \|\Phi_{-1}\|_F^2 = O(np)$. Thus, using again Lemma I.16, we get that, for every $\Delta > 0$ and for all $t' > t$ s.t. $|t - t'| \leq \Delta$,

$$\|\gamma(t) - \gamma(t')\|_2 \leq C_1 p \Delta, \tag{8.84}$$

with probability at least $1 - 2\exp\left(-c_3 \log^2 n\right)$ over X_{-1} and V, where C_1 is an absolute constant independent from the scalings of the problems and from Δ. This means that we can cover T using $\lceil \tau/\Delta \rceil = \left\lceil d \log^2 n/(p\Delta) \right\rceil$ balls with radius $C_1 p \Delta$. Then, setting $\epsilon = C_1 p \Delta$, this implies

$$\mathcal{N}(T, \epsilon) \leq \left\lceil \frac{d \log^2 n}{p} \frac{C_1 p}{\epsilon} \right\rceil = \left\lceil \frac{C_1 d \log^2 n}{\epsilon} \right\rceil. \tag{8.85}$$

Notice that, for $\epsilon \geq \text{diam}(T)$, we simply have $\mathcal{N}(T, \epsilon) = 1$. For smaller values of $\epsilon \leq \text{diam}(T) \leq C_2$, the previous expression reads $\mathcal{N}(T, \epsilon) \leq \frac{C_3 d \log^2 n}{\epsilon}$, where C_3 is another absolute constant set large enough such that $\frac{C_3 d \log^2 n}{\epsilon} \geq 2$, for all $\epsilon \leq C_2$ (this allows to remove the ceiling function), and $C_3 d \log^2 n \geq e C_2$ (which will be used in the next inequality). Then, we have

$$\begin{aligned}
\int_0^\infty \sqrt{\log \mathcal{N}(T, \epsilon)} \mathrm{d}\epsilon &\leq \int_0^{C_2} \sqrt{\log \mathcal{N}(T, \epsilon)} \mathrm{d}\epsilon \\
&\leq \int_0^{C_2} \sqrt{\log\left(\frac{C_3 d \log^2 n}{\epsilon}\right)} \mathrm{d}\epsilon \\
&\leq \int_0^{C_2} \sqrt{\log\left(\frac{eC_2}{\epsilon}\right) + \log\left(\frac{C_3 d \log^2 n}{eC_2}\right)} \mathrm{d}\epsilon \\
&\leq \int_0^{C_2} \sqrt{2\log\left(\frac{eC_2}{\epsilon}\right)} \mathrm{d}\epsilon + \int_0^{C_2} \sqrt{2\log\left(\frac{C_3 d \log^2 n}{eC_2}\right)} \mathrm{d}\epsilon \\
&\leq \sqrt{2} \int_0^{C_2} \log\left(\frac{eC_2}{\epsilon}\right) \mathrm{d}\epsilon + C_2 \sqrt{2\log\left(\frac{C_3 d \log^2 n}{eC_2}\right)} \\
&= \sqrt{2} C_2 + C_4 \log n,
\end{aligned} \tag{8.86}$$

which concludes the argument. □

Proof of Lemma 8.10. Throughout the proof, we condition on the event occurring with probability at least $1 - 2\exp\left(-c_1 \log^2 n\right)$ over V and X given by Lemma I.11. This guarantees that K is invertible and, therefore, that $\Phi^+ = \Phi^\top K^{-1}$, which implies that the LHS of the statement reads $y_i - \varphi(x_i)^\top \hat{\theta}(t) = \varphi(x_i)^\top e^{-\frac{2\Phi^\top \Phi}{n} t} \Phi^+ Y$.

Due to symmetry, we prove the statement for $i = 1$, and the thesis will then be true for all $i \in [n]$ by performing a union bound on the n events. A simple application of the triangle inequality gives

$$\begin{aligned}
&\sup_{t \in [0,\tau]} \left| \varphi(x_1)^\top e^{-\frac{2\Phi^\top \Phi}{n} t} \Phi^+ Y \right| \\
&\leq \sup_{t \in [0,\tau]} \left| \varphi(x_1)^\top e^{-\frac{2\Phi^\top \Phi}{n} t} \left(\Phi^+ Y - \Phi_{-1}^+ Y_{-1} \right) \right| + \sup_{t \in [0,\tau]} \left| \varphi(x_1)^\top e^{-\frac{2\Phi^\top \Phi}{n} t} \Phi_{-1}^+ Y_{-1} \right|.
\end{aligned} \tag{8.87}$$

The first term on the RHS can be bounded as follows

$$\sup_{t\in[0,\tau]}\left|\varphi(x_1)^\top e^{-\frac{2\Phi^\top\Phi}{n}t}\left(\Phi^+Y-\Phi_{-1}^+Y_{-1}\right)\right|\le\|\varphi(x_1)\|_2\left\|\Phi^+Y-\Phi_{-1}^+Y_{-1}\right\|_2=O\left(\log n\right),\tag{8.88}$$

where the last step is a consequence of Lemma 8.14 and $\|\varphi(x_1)\|_2=O\left(\sqrt{p}\right)$ (see (I.147) and (I.148)), and it holds with probability at least $1-2\exp\left(-c_2\log^2 n\right)$ over V and X. To bound the second term on the RHS of (8.87), we first consider the term

$$\sup_{t\in[0,\tau]}\left|\varphi(x_1)^\top e^{-\frac{2\Phi_{-1}^\top\Phi_{-1}}{n}t}\Phi_{-1}^+Y_{-1}\right|\le\sup_{t\in[0,\tau]}\mu_1\left|x_1^\top\gamma(t)\right|+\sup_{t\in[0,\tau]}\left|\tilde\varphi(x_1)^\top\Phi_{-1}^+e^{-\frac{2K_{-1}}{n}t}Y_{-1}\right|,\tag{8.89}$$

where we define the shorthand $\gamma(t)=V^\top e^{-\frac{2\Phi_{-1}^\top\Phi_{-1}}{n}t}\Phi_{-1}^+Y_{-1}$. Using (8.38), the second term reads

$$\sup_{t\in[0,\tau]}\left|\tilde\varphi(x_1)^\top\Phi_{-1}^+e^{-\frac{2K_{-1}}{n}t}Y_{-1}\right|\le\left\|\left(\Phi_{-1}^+\right)^\top\tilde\varphi(x_1)\right\|_2^2\|Y_{-1}\|_2^2=O\left(\log n\right),\tag{8.90}$$

with probability at least $1-2\exp\left(-c_3\log^2 n\right)$ over X and V. For the first term of (8.89), we apply the following decomposition

$$\sup_{t\in[0,\tau]}\left|x_1^\top\gamma(t)\right|\le\sup_{t\in[0,\tau]}\left|\left(x_1-\mathbb{E}_X\left[x_1\right]\right)^\top\gamma(t)\right|+\sup_{t\in[0,\tau]}\left|\mathbb{E}_X\left[x_1\right]^\top\gamma(t)\right|.\tag{8.91}$$

The second term can be easily bounded as follows

$$\sup_{t\in[0,\tau]}\left|\mathbb{E}_X\left[x_1\right]^\top\gamma(t)\right|\le\left\|\mathbb{E}_X\left[x_1\right]\right\|_2\sup_{t\in[0,\tau]}\|\gamma(t)\|_2\le\left\|\mathbb{E}_X\left[x_1\right]\right\|_2\left\|V^\top\Phi_{-1}^+\right\|_{\mathsf{op}}\|Y_{-1}\|_2=O\left(1\right),\tag{8.92}$$

where the last step is a consequence of Lemmas I.7 and I.16, and it holds with probability at least $1-2\exp\left(-c_4\log^2 n\right)$ over X_{-1} and V. For the first term of (8.91), note that the stochastic process $\left(x_1-\mathbb{E}_X\left[x_1\right]\right)^\top\gamma(t)$ is separable, mean-0, and sub-Gaussian, *i.e.*,

$$\left\|\left(x_1-\mathbb{E}[x_1]\right)^\top(u-u')\right\|_{\psi_2}\le\left\|x^\top(u-u')\right\|_{\psi_2}+\|\mathbb{E}[x]\|_2\|u-u'\|_2\le K\|u-u'\|_2,\tag{8.93}$$

for all $u,u'\in T$ (where we use Lemma I.7 and the fact that x is sub-Gaussian by Assumption 8.5). Furthermore, by Lemma 8.17, with probability at least $1-2\exp\left(-c_5\log^2 n\right)$ over X_{-1} and V, we have that, for $t\in[0,\tau]$, $\gamma(t)$ describes a separable set $T\subset\mathbb{R}^d$ such that $\int_0^\infty\sqrt{\log\mathcal{N}(T,\epsilon)}\,\mathsf{d}\epsilon=O\left(\log n\right)$ and $\mathsf{diam}(T)=O\left(1\right)$. Then, conditioning on such event, we can use Dudley's (or chaining tail) inequality (see Theorem 5.29 in van Handel [2016]), which gives

$$\begin{aligned}&\sup_{t\in[0,\tau]}\left|\left(x_1-\mathbb{E}_X\left[x_1\right]\right)^\top\gamma(t)-\left(x_1-\mathbb{E}_X\left[x_1\right]\right)^\top\gamma(0)\right|\\&\qquad\le C_1\left(\int_0^\infty\sqrt{\log\mathcal{N}(T,\epsilon)}\,\mathsf{d}\epsilon+\mathsf{diam}(T)\log n\right)=O\left(\log n\right),\end{aligned}\tag{8.94}$$

with probability at least $1-2\exp\left(-c_6\log^2 n\right)$ over x_1. Furthermore, we have that

$$\begin{aligned}\left|\left(x_1-\mathbb{E}_X\left[x_1\right]\right)^\top\gamma(0)\right|&\le\left|x_1^\top\gamma(0)\right|+\left|\mathbb{E}_X\left[x_1\right]^\top\gamma(0)\right|\\&\le C_2\log n\|\gamma(0)\|_2+\left\|\mathbb{E}_X\left[x_1\right]\right\|_2\|\gamma(0)\|_2\le C_3\|\gamma(0)\|_2\log n=O\left(\log n\right),\end{aligned}\tag{8.95}$$

where the second step holds with probability $1-2\exp\left(-c_7\log^2 n\right)$ over x_1, since it is sub-Gaussian, the third step follows from Lemma I.7, and the last step follows from the same argument used in (8.92), due to Lemma I.16, and holds with probability $1-2\exp\left(-c_8\log^2 n\right)$ over X_{-1} and V.

Plugging (8.94), (8.95), and (8.92) in (8.91) gives $\sup_{t\in[0,\tau]}\left|x_1^\top\gamma(t)\right| = O\left(\log n\right)$ with probability at least $1-2\exp\left(-c_9\log^2 n\right)$ over X and V. Then, this last result, together with (8.90) and (8.89), leads to

$$\sup_{t\in[0,\tau]}\left|\varphi(x_1)^\top e^{-\frac{2\Phi^\top\Phi}{n}t}\Phi_{-1}^+Y_{-1}\right| \le 2\sup_{t\in[0,\tau]}\left|\varphi(x_1)^\top e^{-\frac{2\Phi_{-1}^\top\Phi_{-1}}{n}t}\Phi_{-1}^+Y_{-1}\right| = O\left(\log n\right), \quad (8.96)$$

with probability at least $1-2\exp\left(-c_{10}\log^2 n\right)$, where the first step holds because of Lemma 8.16.

Plugging (8.88) and (8.96) in (8.87), and performing a union bound over all the indices $i\in[n]$, leads to the desired result. We finally remark that this result does not require the condition $\log n = \Theta(\log p)$ in Assumption 8.7. □

Lemma 8.18. *For all $i\in[n]$, we have that*

$$\mathbb{E}_B\left[\left(\varphi(x_i)^\top\tilde\Theta(t)-\varphi(x_i)^\top\tilde\Theta(s)\right)^2\right] \le \Sigma^2\left|t-s\right|\left\|\varphi(x_i)\right\|_2^2. \quad (8.97)$$

Proof. Let us define the matrix $O\in\mathbb{R}^{p\times p}$ as the orthogonal matrix such that $\Phi^\top\Phi = ODO^\top$, with $D\in\mathbb{R}^{p\times p}$ being a p.s.d. diagonal matrix with eigenvalues sorted in non-increasing way (which we indicate as D_k, $k\in[p]$, for simplicity). Then, we have

$$\mathbb{E}_B\left[\left(\varphi(x_i)^\top\tilde\Theta(t)-\varphi(x_i)^\top\tilde\Theta(s)\right)^2\right] = \mathbb{E}_B\left[\left(\varphi(x_i)^\top OO^\top\tilde\Theta(t)-\varphi(x_i)^\top OO^\top\tilde\Theta(s)\right)^2\right]. \quad (8.98)$$

Note that, multiplying both sides of (8.15) by $O^\top$ we get

$$\mathsf{d}\left(O^\top\hat\Theta(t)\right) = -\frac{2D}{n}\left(O^\top\hat\Theta(t)\right)\mathsf{d}t + \frac{2O^\top\Phi^\top Y}{n}\mathsf{d}t + \Sigma\,\mathsf{d}\left(O^\top B(t)\right). \quad (8.99)$$

Since D is diagonal, we have that each coordinate of $O^\top\hat\Theta(t)$ evolves as an independent, one dimensional, OU process (without drift, eventually), since each coordinate of $O^\top B(t)$ evolves as an independent Wiener processes, for rotational invariance of the Gaussian measure. This implies that, for all $s,t\in[0,\tau]$, and for all $k\neq k'$, we have

$$\mathbb{E}_B\left[[O^\top\tilde\Theta(t)]_k[O^\top\tilde\Theta(s)]_{k'}\right] = 0. \quad (8.100)$$

This, together with (8.98), gives

$$\begin{aligned}\mathbb{E}_B\left[\left(\varphi(x_i)^\top\tilde\Theta(t)-\varphi(x_i)^\top\tilde\Theta(s)\right)^2\right] &= \mathbb{E}_B\left[\left(\sum_{k=1}^p[O^\top\varphi(x_i)]_k[O^\top\tilde\Theta(t)]_k-[O^\top\varphi(x_i)]_k[O^\top\tilde\Theta(s)]_k\right)^2\right]\\ &= \sum_{k=1}^p\mathbb{E}_B\left[\left([O^\top\varphi(x_i)]_k[O^\top\tilde\Theta(t)]_k-[O^\top\varphi(x_i)]_k[O^\top\tilde\Theta(s)]_k\right)^2\right]\\ &= \sum_{k=1}^p[O^\top\varphi(x_i)]_k^2\,\mathbb{E}_B\left[\left([O^\top\tilde\Theta(t)]_k-[O^\top\tilde\Theta(s)]_k\right)^2\right].\end{aligned} \quad (8.101)$$

Since $[O^\top \hat{\Theta}(t)]_k$ describes an OU process with drift $\Delta_k = \frac{2D_k}{n}$, standard results on the variance and the auto-correlation of OU processes (see Section 4.4.4 in Gardiner [1985]) give (supposing $\Delta_k \neq 0$)

$$\mathbb{E}_B\left[[O^\top \tilde{\Theta}(t)]_k^2\right] = \frac{\Sigma^2}{2\Delta_k}\left(1 - e^{-2\Delta_k t}\right), \tag{8.102}$$

and

$$\mathbb{E}_B\left[[O^\top \tilde{\Theta}(t)]_k [O^\top \tilde{\Theta}(s)]_k\right] = \frac{\Sigma^2}{2\Delta_k}\left(e^{-\Delta_k |t-s|} - e^{-\Delta_k (t+s)}\right), \tag{8.103}$$

which implies that

$$\mathbb{E}_B\left[\left([O^\top \tilde{\Theta}(t)]_k - [O^\top \tilde{\Theta}(s)]_k\right)^2\right] = \frac{\Sigma^2}{\Delta_k}\left(1 - \frac{e^{-2\Delta_k t} + e^{-2\Delta_k s}}{2} - e^{-\Delta_k |t-s|} + e^{-\Delta_k (t+s)}\right). \tag{8.104}$$

Then, assuming $t \geq s$ without loss of generality, and introducing the shorthand $\delta = t - s$, we have

$$\begin{aligned}
\mathbb{E}_B\left[\left([O^\top \tilde{\Theta}(t)]_k - [O^\top \tilde{\Theta}(s)]_k\right)^2\right] &= \frac{\Sigma^2}{\Delta_k}\left(1 - e^{-2\Delta_k s}\left(\frac{1 + e^{-2\Delta_k \delta} - 2e^{-\Delta_k \delta}}{2}\right) - e^{-\Delta_k \delta}\right) \\
&= \frac{\Sigma^2\left(1 - e^{-\Delta_k \delta}\right)}{\Delta_k}\left(1 - e^{-2\Delta_k s}\left(\frac{1 - e^{-\Delta_k \delta}}{2}\right)\right) \\
&\leq \frac{\Sigma^2\left(1 - e^{-\Delta_k \delta}\right)}{\Delta_k} \\
&\leq \Sigma^2 \delta = \Sigma^2 |t - s|,
\end{aligned} \tag{8.105}$$

where the inequality in the fourth line holds as $1 - e^{-x} \leq x$ for all $x \geq 0$. Note that, when $[O^\top \hat{\Theta}(t)]_k$ does not have a drift term ($\Delta_k = 0$), it is a Wiener process, which implies that $\mathbb{E}_B\left[\left([O^\top \tilde{\Theta}(t)]_k - [O^\top \tilde{\Theta}(s)]_k\right)^2\right] = \Sigma^2 |t - s|$ (see Section 3.8.1 in Gardiner [1985]).

Thus, plugging (8.105) in (8.101) we get

$$\begin{aligned}
\mathbb{E}_B\left[\left(\varphi(x_i)^\top \tilde{\Theta}(t) - \varphi(x_i)^\top \tilde{\Theta}(s)\right)^2\right] &= \sum_{k=1}^{p} [O^\top \varphi(x_i)]_k^2 \, \mathbb{E}_B\left[\left([O^\top \tilde{\Theta}(t)]_k - [O^\top \tilde{\Theta}(s)]_k\right)^2\right] \\
&\leq \sum_{k=1}^{p} [O^\top \varphi(x_i)]_k^2 \, \Sigma^2 \, |t - s| \\
&= \Sigma^2 \, |t - s| \, \|\varphi(x_i)\|_2^2,
\end{aligned} \tag{8.106}$$

which gives the desired result. □

Proof of Lemma 8.11. Consider the auxiliary Wiener process $z_i(t) \in \mathbb{R}$ such that (see Section 3.8.1 of Gardiner [1985])

$$\mathbb{E}_{z_i}\left[\left(z_i(t) - z_i(s)\right)^2\right] = \Sigma^2 \, |t - s| \, \|\varphi(x_i)\|_2^2. \tag{8.107}$$

By the reflection principle in standard Wiener processes (see Proposition 3.7 in Revuz and Yor [2010]), we have

$$\mathbb{P}_{z_i}\left(\sup_{t \in [0,\tau]} z_i(t) \geq a\right) = 2\,\mathbb{P}_{z_i}\left(z_i(\tau) \geq a\right) \leq 4 \exp\left(-\frac{a^2}{\Sigma^2 \tau \, \|\varphi(x_i)\|_2^2}\right). \tag{8.108}$$

By the same argument carried out in (I.147) and (I.148), we have that $\|\varphi(x_i)\|_2^2 = O(p)$ with probability at least $1 - 2\exp(-c_1 p)$ over V. Thus, there exists a constant C for which we can write

$$\Sigma^2 \tau \|\varphi(x_i)\|_2^2 \leq 4C \frac{d\log^6 n}{n^2} \frac{8\log(1/\delta)}{\varepsilon^2} \frac{d\log^2 n}{p} p = 32C \frac{d^2 \log^8 n}{n^2} \frac{\log(1/\delta)}{\varepsilon^2} = O(1), \tag{8.109}$$

where the last step is a consequence of Assumption 8.8. Then, (8.108) reads $\mathbb{P}_{z_i}\left(\sup_{t\in[0,\tau]} z_i(t) \geq a\right) \leq 4\exp(-c_2 a^2)$, which implies

$$\mathbb{E}_{z_i}\left[\sup_{t\in[0,\tau]} |z_i(t)|\right] = \int_0^{+\infty} \mathbb{P}_{z_i}\left(\sup_{t\in[0,\tau]} |z_i(t)| \geq a\right) \mathrm{d}a \leq \int_0^{+\infty} 2\mathbb{P}_{z_i}\left(\sup_{t\in[0,\tau]} z_i(t) \geq a\right) \mathrm{d}a \leq C_1, \tag{8.110}$$

where C_1 is an absolute constant, and the statement holds with probability at least $1 - 2\exp(-c_1 p)$ over V. Note that, by Lemma 8.18, for all $s, t \in [0,\tau]$, we have

$$\mathbb{E}_B\left[\sup_{t\in[0,\tau]} \left|\varphi(x_i)^\top \tilde{\Theta}(t)\right|\right] \leq \Sigma^2 |t-s| \|\varphi(x_i)\|_2^2 = \mathbb{E}_{z_i}\left[(z_i(t) - z_i(s))^2\right]. \tag{8.111}$$

Then, since $\varphi(x_i)^\top \tilde{\Theta}(t)$ and $z_i(t)$ are two mean-0 Gaussian processes, due to Sudakov-Fernique inequality (see Theorem 7.2.11 of Vershynin [2018]), we have that, for all finite subsets $T_0 \subset [0,\tau]$,

$$\mathbb{E}_B\left[\sup_{t\in T_0} \left|\varphi(x_i)^\top \tilde{\Theta}(t)\right|\right] \leq \mathbb{E}_{z_i}\left[\sup_{t\in T_0} |z_i(t)|\right]. \tag{8.112}$$

Differently from Theorem 7.2.11 of Vershynin [2018], we consider the supremum of the absolute value of the processes. The result can in fact be extended to this case by simply looking at both $\varphi(x_i)^\top \tilde{\Theta}(t)$ and $-\varphi(x_i)^\top \tilde{\Theta}(t)$. To extend to the previous equation to the full interval $[0,\tau]$, we have that the Kolmogorov continuity theorem (see Theorem 1.8 in Revuz and Yor [2010]) guarantees that $\varphi(x_i)^\top \tilde{\Theta}(t)$ is almost surely continuous (as its α-moment is bounded by the α-moment of $z_i(t)$ via (8.111) since they are both Gaussian processes). Then, in the compact time interval $[0,\tau]$, we have uniform continuity, *i.e.*, for every $h > 0$, there exists r, such that, for every $t \in [0,\tau]$, $\left|\varphi(x_i)^\top \tilde{\Theta}(t+r) - \varphi(x_i)^\top \tilde{\Theta}(t)\right| < h$, with probability 1 over B. Then, (8.112) can be extended by continuity to

$$\mathbb{E}_B\left[\sup_{t\in[0,\tau]} \left|\varphi(x_i)^\top \tilde{\Theta}(t)\right|\right] \leq \mathbb{E}_{z_i}\left[\sup_{t\in[0,\tau]} |z_i(t)|\right] \leq C_1, \tag{8.113}$$

where the last step is a consequence of (8.110).

Note that, by Lemma 8.18, since we have $\varphi(x_i)^\top \tilde{\Theta}(0) = 0$,

$$\sigma_\tau^2 := \sup_{t\in[0,\tau]} \mathbb{E}_B\left[\left(\varphi(x_i)^\top \tilde{\Theta}(t)\right)^2\right] \leq \Sigma^2 \tau \|\varphi(x_i)\|_2^2 = O(1), \tag{8.114}$$

where in the last step we used (8.109). Then, the Borell-TIS inequality (see Theorem 2.1.1 of Adler and Taylor [2007]) allows us to write, for all $a > 0$,

$$\mathbb{P}_B\left(\sup_{t\in[0,\tau]} \left|\varphi(x_i)^\top \tilde{\Theta}(t)\right| > \mathbb{E}_B\left[\sup_{t\in[0,\tau]} \left|\varphi(x_i)^\top \tilde{\Theta}(t)\right|\right] + a\right) < \exp\left(-\frac{a^2}{2\sigma_\tau^2}\right) \leq 2\exp\left(-c_3 a^2\right). \tag{8.115}$$

Thus, setting $a = \log n$, we get

$$\sup_{t\in[0,\tau]} \left|\varphi(x_i)^\top \tilde{\Theta}(t)\right| \leq \mathbb{E}_B \left[\sup_{t\in[0,\tau]} \left|\varphi(x_i)^\top \tilde{\Theta}(t)\right|\right] + \log n \leq C_1 + \log n = O\left(\log n\right), \quad (8.116)$$

with probability at least $1 - 2\exp\left(-c_3 \log^2 n\right) - 2\exp\left(-c_1 p\right)$ over V and B, where we use (8.113) in the second step. Our argument holds for any choice of i. Then, the thesis holds uniformly on every $i \in [n]$ with probability at least $1 - 2\exp\left(-c_4 \log^2 n\right) - 2n\exp\left(-c_4 p\right)$ over B and V, which provides the desired result. □

8.6.3 Analysis of noise and early stopping

This section contains the proofs of Lemmas 8.12 and 8.13, as well as of our main Theorem 8.9, including a number of auxiliary results required by the argument. We make use of the notation introduced in the previous sections and consider the hyper-parameter choice in (8.13).

Proof of Lemma 8.12. We denote with $A \in \mathbb{R}^{p\times p}$ the p.s.d. matrix such that $A^2 = A^\top A = \mathbb{E}_x\left[\varphi(x)\varphi(x)^\top\right]$ (the definition is well posed as $\mathbb{E}_x\left[\varphi(x)\varphi(x)^\top\right]$ is p.s.d.). Thus,

$$\|A\|_F^2 = \operatorname{tr}\left(A^\top A\right) = \operatorname{tr}\left(\mathbb{E}_x\left[\varphi(x)\varphi(x)^\top\right]\right) = \mathbb{E}_x\left[\operatorname{tr}\left(\varphi(x)\varphi(x)^\top\right)\right] = \mathbb{E}_x\left[\|\varphi(x)\|_2^2\right], \quad (8.117)$$

where we use the linearity of the trace in the third step, and its cyclic property in the fourth step. Furthermore, we can write

$$\|A\|_F^2 = \mathbb{E}_x\left[\|\varphi(x)\|_2^2\right] \leq 2\|\phi(\mathbf{0})\|_2^2 + 2L^2\mathbb{E}_x\left[\|Vx\|_2^2\right] \leq 2\|\phi(\mathbf{0})\|_2^2 + 2L^2\|V\|_{\text{op}}^2\,\mathbb{E}_x\left[\|x\|_2^2\right] = O\left(p\right), \quad (8.118)$$

where we use that ϕ is Lipschitz in the second step, we denote with $\mathbf{0} \in \mathbb{R}^p$ a vector of zeros, and the last step follows from the bound on $\|V\|_{\text{op}}$ given by Lemma I.8, which holds with probability at least $1 - 2\exp\left(-c_1 d\right)$ over V (high probability event over which we will condition until the end of the proof). The introduction of A allows us to rewrite the LHS of (8.24) as $\mathbb{E}_x\left[\left(\varphi(x)^\top\tilde{\Theta}(\tau)\right)^2\right] = \tilde{\Theta}(\tau)^\top\mathbb{E}_x\left[\varphi(x)\varphi(x)^\top\right]\tilde{\Theta}(\tau) = \left\|A\tilde{\Theta}(\tau)\right\|_2^2$, which gives

$$\mathbb{E}_x\left[\left(\varphi(x)^\top\tilde{\Theta}(\tau)\right)^2\right] = \left\|A\tilde{\Theta}(\tau)\right\|_2^2 = \left\|AOO^\top\tilde{\Theta}(\tau)\right\|_2^2 = \left\|AO\tilde{\Delta}\rho\right\|_2^2, \quad (8.119)$$

where $\tilde{\Delta} \in \mathbb{R}^{p\times p}$ is a p.s.d. diagonal matrix s.t. $\tilde{\Delta}^2$ contains $\mathbb{E}_B\left[\left[O^\top\tilde{\Theta}(\tau)\right]_k^2\right]$ in its k-th entry (O is the same matrix considered in Lemma 8.18), and

$$\rho := \tilde{\Delta}^{-1}O^\top\tilde{\Theta}(\tau) \in \mathbb{R}^p \quad (8.120)$$

is a standard Gaussian vector in the probability space of B (since each coordinate of $O^\top\hat{\Theta}(t)$ evolves as an independent, one dimensional, OU process, by the argument following (8.99)). Then, by Theorem 6.3.2 in Vershynin [2018], we have

$$\left\|\left\|AO\tilde{\Delta}\rho\right\|_2 - \left\|AO\tilde{\Delta}\right\|_F\right\|_{\psi_2} \leq C_1\left\|AO\tilde{\Delta}\right\|_{\text{op}} \leq C_1\left\|AO\tilde{\Delta}\right\|_F, \quad (8.121)$$

where the sub-Gaussian norm is intended in the probability space of ρ. This can be turned into the tail inequality (see Proposition 2.5.2 in Vershynin [2018])

$$\mathbb{P}_B\left(\left\|AO\tilde{\Delta}\rho\right\|_2 \geq \left\|AO\tilde{\Delta}\right\|_F + \sqrt{p\Sigma^2\tau}\log n\right) \leq 2\exp\left(-c_2\frac{p\Sigma^2\tau\log^2 n}{\left\|AO\tilde{\Delta}\right\|_F^2}\right). \quad (8.122)$$

Since we have $\left\|AO\tilde{\Delta}\right\|_F^2 \leq \|A\|_F^2 \left\|O\tilde{\Delta}\right\|_{\text{op}}^2 \leq \|A\|_F^2 \left\|\tilde{\Delta}\right\|_{\text{op}}^2 \leq C_2 p\Sigma^2\tau$, where the last step follows from (8.118) (C_2 is an absolute constant) and the argument in (8.105), (8.122) gives

$$\left\|AO\tilde{\Delta}\rho\right\|_2^2 = O\left(p\Sigma^2\tau\log^2 n\right) = O\left(\frac{d^2\log^{10} n}{n^2}\frac{\log(1/\delta)}{\varepsilon^2}\right) = \tilde{O}\left(\frac{d^2}{\varepsilon^2 n^2}\right) = o(1), \quad (8.123)$$

with probability at least $1 - 2\exp\left(-c_2\log^2 n\right)$ over B, where the last step is a consequence of Assumption 8.8. Plugging the last equation in (8.119) provides the desired result. □

Proof of Lemma 8.13. Consider the projector P_Λ on the space spanned by the eigenvectors associated with the d largest eigenvalues of $\Phi^\top\Phi$. Then, let us decompose the LHS of (8.25) as

$$\begin{aligned}
&\mathbb{E}_x\left[\left(\varphi(x)^\top\left(\hat{\theta}(\tau) - \theta^*\right)\right)^2\right] \\
&\leq 2\mathbb{E}_x\left[\left(\varphi(x)^\top P_\Lambda\left(\hat{\theta}(\tau) - \theta^*\right)\right)^2\right] + 2\mathbb{E}_x\left[\left(\varphi(x)^\top P_\Lambda^\perp\left(\hat{\theta}(\tau) - \theta^*\right)\right)^2\right] \\
&\leq 2\mathbb{E}_x\left[\left(\varphi(x)^\top P_\Lambda\left(\hat{\theta}(\tau) - \theta^*\right)\right)^2\right] + 4\mathbb{E}_x\left[\left(\tilde{\varphi}(x)^\top P_\Lambda^\perp\left(\hat{\theta}(\tau) - \theta^*\right)\right)^2\right] \\
&\quad + 4\mu_1^2\mathbb{E}_x\left[\left(x^\top V^\top P_\Lambda^\perp\left(\hat{\theta}(\tau) - \theta^*\right)\right)^2\right].
\end{aligned} \quad (8.124)$$

We now bound the last three terms separately. As usual, we condition on the high probability event given by Lemma I.11, which guarantees that K is invertible and, therefore, $\hat{\theta}(t) = (1 - e^{-\frac{2\Phi^\top\Phi}{n}t})\Phi^+ Y$, with probability at least $1 - 2\exp\left(-c_2\log^2 n\right)$ over V and X, where $\Phi^+ = \Phi^\top K^{-1}$. For the first term of (8.124), exploiting the fact that P_Λ is a projector over an eigenspace of $\Phi^\top\Phi$, we can write

$$P_\Lambda\left(\theta^* - \hat{\theta}(\tau)\right) = P_\Lambda e^{-\frac{2\Phi^\top\Phi}{n}\tau}\Phi^+ Y = P_\Lambda e^{-P_\Lambda\frac{2\Phi^\top\Phi}{n}P_\Lambda\tau}P_\Lambda\Phi^+ Y, \quad (8.125)$$

where $\left\|P_\Lambda e^{-P_\Lambda\frac{2\Phi^\top\Phi}{n}P_\Lambda\tau}P_\Lambda\right\|_{\text{op}} = e^{-\frac{2\lambda_d(K)}{n}\tau} = O\left(e^{-c_3\log^2 n}\right)$, where c_3 is a small enough absolute constant, and the last step is a consequence of Lemma I.11 and $\tau = d\log^2 n/p$, and it holds with probability at least $1 - 2\exp\left(-c_4\log^2 n\right)$ over V and X. Then, using $ne^{-c_3\log^2 n} = O(1)$ we have

$$\left\|P_\Lambda\left(\hat{\theta}(\tau) - \theta^*\right)\right\|_2 \leq \left\|P_\Lambda e^{-P_\Lambda\frac{2\Phi^\top\Phi}{n}P_\Lambda\tau}P_\Lambda\right\|_{\text{op}}\left\|\Phi^+\right\|_{\text{op}}\|Y\|_2 = O\left(\frac{1}{\sqrt{pn}}\right), \quad (8.126)$$

where the second step holds because of Lemma I.11. Then, using (8.118), the first term in (8.124) reads

$$\mathbb{E}_x\left[\left(\varphi(x)^\top P_\Lambda\left(\hat{\theta}(\tau) - \theta^*\right)\right)^2\right] \leq \mathbb{E}_x\left[\|\varphi(x)\|_2^2\right]\left\|P_\Lambda\left(\hat{\theta}(\tau) - \theta^*\right)\right\|_2^2 = O\left(p\frac{1}{pn}\right) = O\left(\frac{1}{n}\right), \quad (8.127)$$

with probability at least $1 - 2\exp\left(-c_5\log^2 n\right)$ over V and X.

For the second term of (8.124), we have

$$\begin{aligned}\mathbb{E}_x\left[\left(\tilde{\varphi}(x)^\top P_\Lambda^\perp\left(\hat{\theta}(\tau)-\theta^*\right)\right)^2\right] &\leq \left\|\mathbb{E}_x\left[\tilde{\varphi}(x)\tilde{\varphi}(x)^\top\right]\right\|_{\mathsf{op}}\left\|\Phi^+\right\|_{\mathsf{op}}^2\|Y\|_2^2\\ &= O\left(\left(\log^4 n+\frac{p\log^3 d}{d^{3/2}}\right)1\frac{1}{p}n\right)\\ &= O\left(\frac{n}{\sqrt{p}}\frac{\log^4 n}{\sqrt{p}}+\frac{n\log^3 d}{d^{3/2}}\right)\\ &= O\left(\frac{d}{n}+\frac{n\log^3 d}{d^{3/2}}\right)=o(1).\end{aligned} \tag{8.128}$$

Here, the second line follows from Lemmas I.21 and I.11, and it holds with probability at least $1-2\exp\left(-c_6\log^2 n\right)$ over X and V.

For the third term of (8.124), since x is distributed according to $\mathcal{P}_X$, it is sub-Gaussian and $\|x\|_{\psi_2}=O(1)$. Then, we can bound its second moment (see Vershynin [2018], Proposition 2.5.2) as follows

$$\begin{aligned}\mathbb{E}_x\left[\left(x^\top V^\top P_\Lambda^\perp\left(\hat{\theta}(\tau)-\theta^*\right)\right)^2\right] &\leq C_1\left\|V^\top P_\Lambda^\perp\left(\hat{\theta}(\tau)-\theta^*\right)\right\|_2^2\\ &= C_1\left\|V^\top P_\Lambda^\perp\Phi^+e^{-\frac{2K}{n}\tau}Y\right\|_2^2\\ &\leq C_1\left\|V^\top P_\Lambda^\perp\Phi^+\right\|_{\mathsf{op}}^2\|Y\|_2^2\\ &= O\left(\frac{d}{n}+\frac{n^2\log^6 d}{d^3}\right)\\ &= O\left(\frac{d}{n}+\frac{n\log^3 d}{d^{3/2}}\right)=o(1),\end{aligned} \tag{8.129}$$

where C_1 is an absolute constant and the fourth line holds with probability $1-2\exp\left(-c_7\log^2 n\right)$ over X and V, because of Lemma I.20. Plugging (8.127), (8.128), and (8.129) in (8.124) provides the desired result. □

We would like to remark that a key aspect of the previous argument is the existence of a *spectral gap* between $\lambda_d(K)$ and $\lambda_{d+1}(K)$. This result, proved in Lemma I.11, critically uses that $\lambda_{\min}\left(X^\top X\right)=\Omega(n)$, which in turn relies on our assumption $\lambda_{\min}\left(\mathbb{E}\left[x^\top x\right]\right)=\Omega(1)$, *i.e.*, that the data covariance is well-conditioned. The spectral gap is what allows us to define a proper early stopping time, which implicitly connects to the spectrum of K via the previous argument. We note that, if $\lambda_{\min}\left(\mathbb{E}\left[x^\top x\right]\right)=o(1)$, then K might lose its spectral gap, a new argument would be required to set the early stopping time, and we eventually expect the final bound on the excess population risk to depend on the condition number of the covariance (as it does also in prior related work [Varshney et al., 2022, Liu et al., 2023b]).

When discussing Figure 8.3, we have commented on the regularizing effect of DP-GD which can be connected to the introduction of a ridge penalty. In fact, adding Gaussian noise and early stopping does not allow DP-GD to interpolate the training samples in a similar way to ℓ_2 regularization. The connection between ℓ_2 regularization and early stopping can be made quantitative by inspecting the argument of the previous Lemma 8.13: there, we show that the gradient flow at early stopping is close to the gradient flow at convergence when looking only at the sub-space S_Λ spanned by the eigenvectors of $\Phi^\top\Phi$ associated to the top-d eigenvalues

(see (8.126)), while the same is not true in the orthogonal sub-space $S_\Lambda^\perp$. When we look at the ridge solution

$$\theta_\lambda^* = \left(\Phi^\top \Phi + \lambda I\right)^{-1} \Phi^\top Y, \tag{8.130}$$

if λ is chosen such that $\lambda_{d+1}(\Phi^\top\Phi) \ll \lambda \ll \lambda_d(\Phi^\top\Phi)$, we have a similar effect: in the sub-space S_Λ, θ_λ^* is approximately the same as the un-regularized θ^*, while we still avoid over-fitting due to the differences in the orthogonal sub-space $S_\Lambda^\perp$.

Proof of Theorem 8.9. First, note that since our model has a quadratic loss, following the first two equations in Assumption 8.8 we can apply Proposition 8.4. Then, the hyper-parameter Σ in (8.13) is sufficiently large to guarantee that the solution $\Theta(\tau)$ of the SDE (8.8) at time τ is (ε, δ)-differentially private. Let us introduce the shorthand $R(x) = \varphi(x)^\top \left(\hat{\Theta}(\tau) - \theta^*\right)$, and denote the generalization error of $\hat{\Theta}(\tau)$ and θ^* by $\hat{\mathcal{R}}$ and $\mathcal{R}^*$, respectively. Then, we have that

$$\begin{aligned}
\left(\hat{\mathcal{R}} - \mathcal{R}^*\right)^2 &= \left(\mathbb{E}_{(x,y)\sim\mathcal{P}_{XY}}\left[\left(\varphi(x)^\top\hat{\Theta}(\tau) - y\right)^2\right] - \mathbb{E}_{(x,y)\sim\mathcal{P}_{XY}}\left[\left(\varphi(x)^\top\theta^* - y\right)^2\right]\right)^2 \\
&= \left(\mathbb{E}_{(x,y)\sim\mathcal{P}_{XY}}\left[R(x)\left(\left(\varphi(x)^\top\hat{\Theta}(\tau) - y\right) + \left(\varphi(x)^\top\theta^* - y\right)\right)\right]\right)^2 \\
&\leq \mathbb{E}_x\left[R(x)^2\right]\mathbb{E}_{(x,y)\sim\mathcal{P}_{XY}}\left[\left(\varphi(x)^\top\hat{\Theta}(\tau) + \varphi(x)^\top\theta^* - 2y\right)^2\right] \\
&= \mathbb{E}_x\left[R(x)^2\right]\mathbb{E}_{(x,y)\sim\mathcal{P}_{XY}}\left[\left(R(x) + 2\varphi(x)^\top\theta^* - 2y\right)^2\right] \\
&\leq 2\mathbb{E}_x\left[R(x)^2\right]^2 + 2\mathbb{E}_x\left[R(x)^2\right]\mathbb{E}_{(x,y)\sim\mathcal{P}_{XY}}\left[\left(2\varphi(x)^\top\theta^* - 2y\right)^2\right] \\
&\leq 2\mathbb{E}_x\left[R(x)^2\right]^2 + 16\mathbb{E}_x\left[R(x)^2\right]\mathbb{E}_x\left[\left(\varphi(x)^\top\theta^*\right)^2\right] + 16\mathbb{E}_x\left[R(x)^2\right]\mathbb{E}_y\left[y^2\right],
\end{aligned} \tag{8.131}$$

where we factorize the difference of two squares in the second line and use Cauchy-Schwartz inequality in the third line. By Lemmas 8.12 and 8.13, we have that

$$\begin{aligned}
\mathbb{E}_x\left[R(x)^2\right] &= \mathbb{E}_x\left[\left(\varphi(x)^\top\left(\hat{\theta}(\tau) + \tilde{\Theta}(\tau) - \theta^*\right)\right)^2\right] \\
&\leq 2\mathbb{E}_x\left[\left(\varphi(x)^\top\tilde{\Theta}(\tau)\right)^2\right] + 2\mathbb{E}_x\left[\left(\varphi(x)^\top\left(\hat{\theta}(\tau) - \theta^*\right)\right)^2\right] \\
&= O\left(\frac{d^2\log^{10} n}{n^2}\frac{\log(1/\delta)}{\varepsilon^2} + \frac{d}{n} + \frac{n\log^3 d}{d^{3/2}}\right) = o\left(1\right),
\end{aligned} \tag{8.132}$$

with probability at least $1 - 2\exp\left(-c_1\log^2 n\right)$ over X, V and B. Furthermore, by Lemma I.22, we also have that $\mathbb{E}_x\left[\left(\varphi(x)^\top\theta^*\right)^2\right] = O\left(1\right)$ with probability at least $1 - 2\exp\left(-c_2\log^2 n\right)$ over X and V. Since also $\mathbb{E}_y\left[y^2\right] = O\left(1\right)$ by Assumption 8.5, plugging (8.132) in (8.131) gives

$$\left(\hat{\mathcal{R}} - \mathcal{R}^*\right)^2 = O\left(\frac{d^2\log^{10} n}{n^2}\frac{\log(1/\delta)}{\varepsilon^2} + \frac{d}{n} + \frac{n\log^3 d}{d^{3/2}}\right) = o(1), \tag{8.133}$$

with probability at least $1 - 2\exp\left(-c_3 \log^2 n\right)$ over X, V and B. Then, since $\sqrt{a+b+c} \leq \sqrt{a} + \sqrt{b} + \sqrt{c}$, we have that, with this same probability,

$$\begin{aligned} \left|\hat{\mathcal{R}} - \mathcal{R}^*\right| &= O\left(\frac{d}{n\varepsilon}\,\log^5 n\sqrt{\log(1/\delta)} + \sqrt{\frac{d}{n}} + \sqrt{\frac{n\log^3 d}{d^{3/2}}}\right) \\ &= \tilde{O}\left(\frac{d}{n\varepsilon} + \sqrt{\frac{d}{n}} + \sqrt{\frac{n}{d^{3/2}}}\right) = o(1). \end{aligned} \tag{8.134}$$

To conclude the argument, we can now use Lemmas 8.10 and 8.11, which guarantee that, jointly for all $i \in [n]$,

$$\sup_{t\in[0,\tau]} \left(\left|\varphi(x_i)^\top \hat{\Theta}(t)\right| + \left|\varphi(x_i)^\top \hat{\theta}(t) - y_i\right|\right) \leq C_1 \log n, \tag{8.135}$$

with probability at least $1 - 2\exp(-c_4 \log^2 n)$ over X, V, and B, for some absolute constant C_1. Furthermore, with probability at least $1 - 2\exp(-c_5 p)$ over V, $\|\varphi(x_i)\|_2 \leq C_2\sqrt{p}$ for some absolute constant C_2, jointly for every $i \in [n]$ (see the argument carried out in (I.147) and (I.148)). Thus, the previous equation gives

$$\begin{aligned} \sup_{t\in[0,\tau]} \left(\left|\varphi(x_i)^\top \hat{\Theta}(t)\right| + \left|\varphi(x_i)^\top \hat{\theta}(t) - y_i\right|\right) &\leq C_1 \log n \leq \frac{C_2\sqrt{p}}{\|\varphi(x_i)\|_2} C_1 \log n \\ &\leq \frac{\sqrt{p}\log^2 n}{2\,\|\varphi(x_i)\|_2} = \frac{C_{\mathsf{clip}}}{2\,\|\varphi(x_i)\|_2}, \end{aligned} \tag{8.136}$$

with probability at least $1 - 2\exp(-c_6 \log^2 n)$ over X, V, and B, where c_6 may depend on C_1 and C_2 due to the step in the second line. Hence, due to the argument between (8.16) and (8.20), we have that, with the same probability,

$$\Theta(\tau) = \hat{\Theta}(\tau), \tag{8.137}$$

which in turn implies $\hat{\mathcal{R}} = \mathcal{R}$, where $\mathcal{R}$ is the generalization error of $\Theta(\tau)$. This, together with (8.134), concludes the proof. □

8.7 Concluding remarks

In this chapter, we show that privacy can be obtained for free, i.e., the excess population risk is negligible even when $\varepsilon = o(1)$, in sufficiently over-parameterized random feature models. This proves that over-parameterization is not at odds with private learning.

We focus on the regime where the number of samples n scales with the input dimension d as $d \ll n \ll d^{3/2}$, but we believe that our approach can be applied in much wider generality. In fact, we crucially leverage the fact that the test loss of the GD solution θ^* does not decrease with respect the case $n = \Theta(d)$. This in turn means that there is a surplus of samples that can be used to learn privately. A similar plateau in the test loss has been demonstrated for kernel ridge regression by a recent line of work [Ghorbani et al., 2021, Mei et al., 2021, Hu et al., 2024] tackling increasingly general settings, and showing privacy for free here is an exciting direction for future research.

CHAPTER 9

High-Dimensional Private Linear Regression with Optimal Rates

Differentially private (DP) linear regression has received significant attention in the recent theoretical literature, with several approaches proposed to improve error rates. This chapter considers the popular high-dimensional regime with random data, where the number of training samples n and the input dimension d grow at a proportional rate $d/n \to \gamma$, and it studies a family of one-pass DP gradient descent (DP-GD) algorithms satisfying $\rho^2/2$ zero concentrated DP. In this setting, we establish a deterministic equivalent for the DP-GD trajectory in terms of a system of ordinary differential equations. This allows to analyze the effect of gradient clipping constants that are smaller than the typical norm of the per-sample gradients — a setup shown to improve performance in practice. For well-conditioned data, we show that DP-GD, upon properly choosing clipping constant and learning rate, achieves the non-asymptotic risk of $O(\gamma + \gamma^2/\rho^2)$, and we establish that this rate is minimax optimal. Then, we consider the ill-conditioned case where the data covariance spectrum follows a power-law distribution, and we show that the risk displays a power-like scaling law in γ, highlighting the change in the exponent as a function of the privacy parameter ρ. Overall, our analysis demonstrates the benefits of practical algorithmic design choices, including aggressive gradient clipping and decaying learning rate schedules.

This chapter is based on the work under review: *High-dimensional private linear regression with optimal rates*.

9.1 Introduction

Differential privacy (DP) [Dwork et al., 2006] has consolidated as the standard framework for privacy guarantees and data protection in machine learning. This has motivated an extensive research effort on theoretical problems connected to statistics and optimization [Chaudhuri et al., 2011, Bassily et al., 2014, 2019, Kamath et al., 2019, 2020, Cai et al., 2021, 2025], on its implementation in training large scale deep learning architectures [Li et al., 2022b, De et al., 2022, McKenna et al., 2025], and on its adoption for the release of public data [Hawes, 2020, Desfontaines, 2021]. This framework quantifies privacy through numerical parameters (e.g. ρ in Definition 9.1), which upper bound the impact a single data point can have on the output of the algorithm. Then, guaranteeing a fixed level of DP imposes specific constraints

on the learning algorithm, which come with a performance cost: the more stringent the privacy requirements (corresponding to smaller values of ρ), the higher is such cost. For example, in first-order optimization, these constraints are typically implemented through *clipping* the per-sample gradients if the loss is not Lipschitz, limiting the number of data access via *early stopping*, and adding independent *noise* at every update [Song et al., 2013, Bassily et al., 2014, Abadi et al., 2016].

Understanding the cost of privacy in different learning settings and how to achieve optimal performance has been at the center of a rich line of work spanning over a decade, see e.g. Chaudhuri et al. [2011], Bassily et al. [2014, 2019], Song et al. [2021b], Cai et al. [2021], Varshney et al. [2022]. On the one hand, this involves establishing statistical lower bounds on the expected risk any algorithm must incur when learning with differential privacy; on the other hand, this involves defining efficient algorithms that provably achieve this risk. A notorious example is the problem of *DP linear regression*, where one is given n input-label pairs (x_i, y_i), sampled i.i.d. from a joint distribution P_{XY} such that $y_i = x_i^\top \theta^* + z_i$, where $\theta^* \in \mathbb{R}^d$, $x_i \in \mathbb{R}^d$ has mean-0 and covariance $\Sigma \in \mathbb{R}^{d\times d}$, and $z_i \in \mathbb{R}$ is independent label noise with mean-0 and variance ζ^2. Informally, the goal is to provide an algorithm $\mathcal{M}$ that, given the training set, outputs a parameter θ^p guaranteeing a required privacy budget and minimizing the *test risk*:

$$\mathcal{R}(\mathcal{M}) = \frac{1}{2}\mathbb{E}_{(x,y)\sim P_{XY}}\left[\left(x^\top\theta^p - y\right)^2\right] - \frac{\zeta^2}{2} = \frac{\left\|\Sigma^{1/2}\left(\theta^p - \theta^*\right)\right\|_2^2}{2}. \tag{9.1}$$

On the side of statistical lower bounds, Cai et al. [2021] study the maximum risk over a set $\mathcal{B}$ of possible θ^*, i.e. $\sup_{\theta^*\in\mathcal{B}} \mathcal{R}(\mathcal{M})$, and Theorem 4.1 therein shows that, if $\mathbb{M}$ represents the set of all $\rho^2/2$ zCDP algorithms[1] and $\mathcal{B}$ is the unit Euclidean ball, the *minimax rates* of this problem satisfy

$$\inf_{\mathcal{M}\in\mathbb{M}} \sup_{\theta^*\in\mathcal{B}} \mathbb{E}\left[\mathcal{R}(\mathcal{M})\right] = \Omega\left(\frac{d}{n} + \frac{d^2}{\rho^2 n^2 \log n}\right), \tag{9.2}$$

where the expectation is taken with respect to the randomness of the data and of the private algorithm.

On the algorithmic side, a number of recent papers gives efficient algorithms with theoretical guarantees on the test risk [Wang, 2018, Cai et al., 2021, Milionis et al., 2022, Varshney et al., 2022, Liu et al., 2023b, Brown et al., 2024b]. In particular, Varshney et al. [2022], Liu et al. [2023b], Brown et al. [2024b] propose gradient-based algorithms for which $n = \tilde{\Omega}(d)$ samples are sufficient to achieve non trivial performance when ρ is of constant order (here, d denotes the input dimension and the $\sim$ hides logarithmic factors). For example, Brown et al. [2024b] show that a $\rho^2/2$ zero concentrated DP (zCDP) gradient descent (DP-GD) algorithm $\mathcal{M}$ achieves an error of

$$\mathcal{R}(\mathcal{M}) = O\left(\frac{d}{n} + \frac{d^2}{n^2\rho^2}\operatorname{poly}\log n\right). \tag{9.3}$$

We note that Varshney et al. [2022], Liu et al. [2023b], Brown et al. [2024b] consider gradient updates on the square loss $(x_i^\top\theta - y_i)^2$. As this loss is non Lipschitz, in order to bound the *sensitivity* of the parameter updates with respect to any single training data point, the gradient is *clipped*. The clipping then ensures that the ℓ_2 norm of the gradient does not exceed a custom hyper-parameter C_{clip}, which in turn defines the amount of private noise introduced at every iteration (see, e.g., Algorithm 1 in Brown et al. [2024b]). Notably, a common approach

[1] Cai et al. [2021] give their result in terms of (ε, δ)-DP, and (9.2) can be obtained through the conversion from zCDP to approximated DP stated in Proposition J.3.

in these works is to set the *clipping constant* C_{clip} to be sufficiently large so that, with high probability, gradient clipping effectively *does not take place* throughout the algorithm. This simplifies the analysis, as the optimization becomes more easily comparable to a quadratic problem with noisy gradient updates, but it introduces a larger amount of private noise at every iteration, which gives the additional poly-logarithmic factor in (9.3). In fact, the practical benefit of avoiding gradient clipping is unclear, and it is empirically pointed out in Brown et al. [2024b] that the lowest error occurs under *significant clipping*, see Figure 3 therein. This last conclusion is also in agreement with experimental evidence for deep learning optimization [Kurakin et al., 2022, Li et al., 2022b, De et al., 2022], which supports setting a sufficiently small C_{clip}, with the caveat of properly rescaling either the learning rate or the number of training iterations. At the same time, other empirical work suggests an inherent trade-off in the choice of C_{clip} [Amin et al., 2019, Andrew et al., 2021], due to the bias induced by small clipping constants [Chen et al., 2020a, Song et al., 2021b]. In short, this picture remains poorly understood, also on the theoretical side: to the best of our knowledge, there is no analysis of the dynamics of DP-GD when C_{clip} is of the order of the per-sample gradients, nor precise results quantifying the potential benefits of small clipping constants. Hence, practitioners are left without clear guidance, highlighting the need for a theoretical understanding of the impact of this hyper-parameter.

In this chapter, we consider a high-dimensional setting with random data, where the number of training samples n grows proportionally with the number of input dimensions d so that $d/n \to \gamma \in (0, +\infty)$, with constant-order privacy parameter ρ. This *proportional regime* has been object of extensive study in the context of high-dimensional statistics and machine learning [Hastie et al., 2022, Sur and Candès, 2019, Mei and Montanari, 2022], owing its popularity also to the fact that, in modern data science, the feature dimension of the dataset is often comparable to the number of samples. We note that, in this regime, *(i)* the lower bound in (9.2) does not capture the cost of privacy, since the second term depending on ρ is always negligible with respect to the first term; and *(ii)* the upper bound in (9.3) gives vacuous guarantees, due to the extra poly-logarithmic factor. Non-trivial test risk guarantees in the proportional regime are established by Dwork et al. [2025] for the solution of minimization problems with output and objective perturbation. However, the analysis is limited to isotropic data covariance and it does not characterize how the risk behaves as a function of γ. As concerns gradient-based methods, Dwork et al. [2025] provide a characterization in terms of dynamical mean-field theory (DMFT) equations [Celentano et al., 2021, Gerbelot et al., 2024, Han, 2024]. However, this characterization is hard to interpret, as it does not provide an explicit trajectory or rate, but rather a self consistent, non-Markovian dynamical system (see, e.g., Theorem 7.1 in Dwork et al. [2025]). In contrast, this chapter considers a DP-GD algorithm, deriving the non-asymptotic behavior of its test risk as $\gamma \to 0$. Our analysis allows to characterize the role of the hyper-parameters of the algorithm, such as the learning rate and the clipping constant, also in the regime where C_{clip} is smaller than the typical norm of the per-sample gradients. Then, upon optimally choosing these hyper-parameters, we show that DP-GD achieves the rate $O(\gamma + \gamma^2/\rho^2)$, which we prove to be minimax optimal for algorithms respecting $\rho^2/2$-zCDP.

9.1.1 Main results and contributions

More precisely, we focus on a one-pass DP-GD[2] algorithm (see Algorithm 9.1), with privacy guarantees on its final output expressed in terms of zero concentrated DP (Proposition 9.2). Our main technical contributions are summarized below.

1. Following recent progress in high-dimensional optimization [Paquette et al., 2024a, Collins-Woodfin et al., 2024a,b, Marshall et al., 2025], we show that, as $d, n \to \infty$ such that $d/n \to \gamma \in (0, +\infty)$, the test risk of DP-GD is well described by the solution of a deterministic family of d coupled ordinary differential equations (ODEs), see Theorem 9.6. Importantly, this approach allows to study the setting where the clipping constant C_{clip} is smaller than the typical size of the per-sample gradients, characterizing the error rates of DP-GD via the non-asymptotic behavior of the system of ODEs as $\gamma \to 0$.

2. Next, we focus on the well-conditioned case, i.e., where the eigenvalues of the data covariance do not scale with γ. In this setting, we demonstrate the benefits of using small clipping constants and decaying the learning rate, see Proposition 9.9. Then, in Theorem 9.11, we show that, through a harmonically decreasing learning rate schedule and a properly set clipping constant, Algorithm 9.1 achieves a test risk of

$$\mathcal{R}(\mathcal{M}) = O\left(\gamma + \frac{\gamma^2}{\rho^2}\right), \tag{9.4}$$

   both in high probability and in expectation. These are the *optimal rates* for this task: we give a matching minimax lower bound in Theorem 9.12. More precisely, we show that, if $\mathbb{M}$ represents the set of all $\rho^2/2$ zCDP algorithms and $\mathcal{B}$ is the unit Euclidean ball, then

$$\inf_{\mathcal{M}\in\mathbb{M}} \sup_{\theta^*\in\mathcal{B}} \mathbb{E}\left[\mathcal{R}(\mathcal{M})\right] = \Omega\left(\gamma + \frac{\gamma^2}{\rho^2}\right). \tag{9.5}$$

3. Finally, we study the ill-conditioned case, i.e., where the covariance spectrum follows a power-law distribution. In this setting, we show that the test risk also follows a power law of the form γ^h, see Theorem 9.14. More precisely, via a tight analysis of the system of coupled ODEs, we characterize the exponent h in terms of the privacy level ρ, the learning rate schedule and the decays of the covariance spectrum and of the target regression coefficients θ^*.

Our results are illustrated via experiments on synthetically generated data as well as on standard datasets (MNIST, California housing prices). These showcase *(i)* the accuracy of the deterministic ODEs (Figure 9.1), *(ii)* the predictive behavior of our proposed scaling laws (Figures 9.2 and 9.3), *(iii)* the benefits of taking a small clipping constant and the connection between clipping and an appropriate choice for the learning rate (Figures 9.4, 9.5, 9.6), and *(iv)* the impact on performance of decaying the learning rate during training (Figures 9.7 and 9.8). We finally remark that our analysis is fueled by a methodology that is both innovative compared to earlier work in the DP optimization literature [Varshney et al., 2022, Liu et al., 2023b, Brown et al., 2024b] and rather general, thus laying the foundations for the study of DP algorithms in the high-dimensional setting.

[2]We use the term GD rather than stochastic gradient descent (SGD) to emphasize that our algorithm does not incorporate techniques such as privacy amplification via sub-sampling or shuffling.

9.2 Related work

DP optimization. Beyond the discussion on related literature on DP optimization in Chapter 8, in DP linear regression Wang [2018], Milionis et al. [2022] require $n = \tilde{\Omega}(d^{3/2})$ samples, where the privacy budget is assumed of constant order, while Anderson et al. [2025] require $n = \Omega(d^2)$ samples to also achieve robustness under a corrupted data model. DP-GD with adaptive clipping is shown to achieve a sample complexity of $n = \tilde{\Omega}(d)$ in Varshney et al. [2022]. This is "nearly optimal", in the sense that it matches, up to logarithmic factors, the minimax lower bound in Cai et al. [2021]. Similar nearly optimal rates were previously obtained by Liu et al. [2022] (however with a computationally inefficient method), and more recently by Liu et al. [2023b] and Brown et al. [2024b]. Notably, in the proportional regime $n = \Theta(d)$, if the privacy budget is of constant order ($\varepsilon/\sqrt{\ln(1/\delta)} = \Theta(1)$ in Varshney et al. [2022], Liu et al. [2023b], or $\rho = \Theta(1)$ in Brown et al. [2024b]), these bounds on the test risk diverge logarithmically either in n [Varshney et al., 2022] or in the failure probability [Liu et al., 2023b, Brown et al., 2024b][3]. Precise constant-order values for the test risk in the proportional regime have been provided by the recent work of Dwork et al. [2025] for private algorithms based on output and objective perturbation in the case of isotropic data covariance, without focusing on the non-asymptotic rates in γ.

Minimax lower bounds. Minimax lower bounds in statistical estimation are classically derived via tools such as Le Cam's method, Fano's inequality, and Assouad lemma [Tsybakov, 2009, Wainwright, 2019]. In a linear model with Gaussian data, a convenient way to prove minimax lower bounds is to choose a prior on θ^* and compute the corresponding Bayes risk. Under DP, early minimax lower bounds were developed by Barber and Duchi [2014], with Acharya et al. [2021] later formulating private analogues of Le Cam, Fano and Assouad. Fingerprinting codes and tracing attack arguments were introduced by Bun et al. [2014], Dwork et al. [2015], and those were used to provide statistical lower bounds for high-dimensional mean estimation and linear regression by Cai et al. [2021]. More recently, tracing attacks have been generalized to the score attack technique discussed in Cai et al. [2025], which allows to treat more general statistical models with a well defined score statistic. In the context of local DP, where each sample is privatized prior to observation, the statistical cost of privacy has been studied in Duchi et al. [2013, 2018], Rohde and Steinberger [2020], Duchi and Ruan [2024].

Learning in high dimensions. The setting where the input dimension (or model size) d scales with the number of training samples n gained popularity for its ability to explain several empirical phenomena observed in modern data analysis and deep learning [Hastie et al., 2022, Belkin et al., 2019a, Bartlett et al., 2020, 2021]. Early investigations of this regime appear in the work of Huber [1973] in the context of robust regression. More recently, Hastie et al. [2022] studied linear regression in the proportional asymptotic regime $d/n \to \gamma$, analyzing the risk of ridge(-less) regression in both the under-parameterized ($\gamma < 1$) and over-parameterized ($\gamma > 1$) regime. The proportional asymptotic in ridge regression had been previously considered by Dicker [2016], Dobriban and Wager [2018], while Cheng and Montanari [2024] went beyond this regime establishing non-asymptotic guarantees with multiplicative error bounds. Sur and Candès [2019] studied the maximum likelihood estimator in logistic regression in the proportional regime, highlighting its qualitative differences with respect to

[3] More precisely, in Liu et al. [2023b] the number of samples would not be sufficient to achieve Eq. (4) with high probability.

the case where $d = o(n)$; Deng et al. [2022], Montanari et al. [2025] studied logistic models, considering the effects of over-parameterization and benign overfitting. Another line of work, closer to our technical approach, analyzes the behavior of one-pass GD algorithms in terms of high-dimensional stochastic differential equations or coupled ODEs [Paquette et al., 2022, 2024a, Collins-Woodfin et al., 2024a, Marshall et al., 2025]. This strategy gives a deterministic equivalent for the gradient dynamics, which in turn leads to insights on optimization stability and the role of stochastic batching. Another common analytical tool to analyze the high-dimensional regime is the convex Gaussian minimax theorem [Stojnic, 2013, Thrampoulidis et al., 2015], which constitutes the main technical tool in Dwork et al. [2025].

Gradient clipping. In the context of private optimization with a non-Lipschitz loss, the role of clipping and the magnitude of the corresponding clipping constant C_{clip} has attracted attention due to its nuanced implications: while a small C_{clip} significantly affects the gradients, larger values force the addition of more private noise, suggesting that the choice of C_{clip} induces a bias-variance trade-off [McMahan et al., 2018, Amin et al., 2019, Andrew et al., 2021, Das et al., 2023, Brown et al., 2024b]. Prior work has argued that the bias induced by small clipping constants can prevent convergence [Amin et al., 2019, Chen et al., 2020a, Song et al., 2021b], which motivates an adaptive selection of C_{clip} based on (private) statistics of the magnitude of the gradients [Abadi et al., 2016, Andrew et al., 2021, Pichapati et al., 2019, Golatkar et al., 2022]. Recent experimental studies have given evidence that the best performance is achieved with a sufficiently small C_{clip} [Kurakin et al., 2022, Li et al., 2022b, De et al., 2022], but it has also been shown that overly-aggressive clipping can be damaging in the context of model calibration [Bu et al., 2023, Brown et al., 2024b]. Theoretical insights on the benefits of small clipping constants have been provided in Das et al. [2023], Chen et al. [2020a], with Chen et al. [2020a] proving optimization guarantees when the gradient distribution is sufficiently symmetric and Das et al. [2023] considering the setting where the Lipschitz constant of the loss is sample-dependent. Recent work on DP linear regression [Varshney et al., 2022, Liu et al., 2023b, Brown et al., 2024b] shares the common feature of setting the (possibly adaptive) C_{clip} a poly-logarithmic factor larger than the expected norm of the per-sample gradient. This ensures that, with high probability, at most a few gradients are clipped during training, and we note that these logarithmic factors are strongly tied to the consequent logarithmic divergence of the test risk guarantees discussed in the previous paragraph. Marshall et al. [2025] provide a precise analysis of a version of clipped-GD, although they do not focus on privacy and, therefore, they do not add private noise. In a parallel work, Lin et al. [2025b] add private noise but do not consider clipping, formulating instead a notion of privacy in terms of a diffusion surrogate of the algorithm which is different from a DP guarantee on the algorithm itself.

Scaling laws. Large scale experiments on modern AI systems revealed how the risk improves according to a power-law in the model size and number of training data [Kaplan et al., 2020, Hoffmann et al., 2022]. This evidence sparked a recent line of work establishing provable scaling laws in simplified settings such as linear regression and shallow neural networks [Paquette et al., 2024b, Defilippis et al., 2024, Lin et al., 2024, 2025a, Ren et al., 2025, Ferbach et al., 2025, Wu et al., 2026]. These results generally achieve power-law type scaling laws assuming the data covariance spectrum itself follows a power-law distribution. Most similar to our setting (see Assumption 9.13) is the work by Collins-Woodfin et al. [2024b], which establishes convergence rates for the risk in the context of one-pass GD with adaptive learning rates.

9.3 Preliminaries

Notation. Given a vector v, we denote by $\|v\|_2$ its Euclidean norm. Given a matrix A, we denote by $\mathrm{tr}(A)$ and $\|A\|_{\mathsf{op}}$ its trace and operator (spectral) norm. Given a symmetric matrix A, we denote by $\lambda_{\min}(A)$ ($\lambda_{\max}(A)$) its smallest (largest) eigenvalue. The complexity notations $\Omega(\cdot)$, $O(\cdot)$, $\omega(\cdot)$, $o(\cdot)$ and $\Theta(\cdot)$ are understood for large data size n and input dimension d, while the notation $\Omega_\gamma(\cdot), O_\gamma(\cdot), \Theta_\gamma(\cdot), \omega_\gamma(\cdot), o_\gamma(\cdot)$ is intended for sufficiently small γ. We indicate with $C > 0$ a numerical constant independent of n and d, whose value may change from line to line, and we say that an event holds with overwhelming probability if it holds with probability at least $1 - e^{-\omega(\log d)}$.

Linear regression. Let (X, Y) be a labeled training dataset, where $X = [x_1, \dots, x_n]^\top \in \mathbb{R}^{n\times d}$ contains the training data on its rows and $Y = [y_1, \dots, y_n]^\top \in \mathbb{R}^n$ contains the corresponding labels, such that the input-label pairs are sampled i.i.d. from a joint distribution P_{XY}. We consider a *linear regression* model, where the labels are defined as

$$y_i = x_i^\top \theta^* + z_i, \tag{9.6}$$

where $\theta^* \in \mathbb{R}^d$, x_i has mean-0 and covariance $\Sigma \in \mathbb{R}^{d\times d}$, and z_i is independent label noise with mean 0 and variance ζ^2. We use the shorthand (ω_i, λ_i) to denote an eigenvector-eigenvalue pair of the data covariance Σ, and let $\lambda_{\max} = \lambda_{\max}(\Sigma), \lambda_{\min} = \lambda_{\min}(\Sigma)$ denote its largest and smallest eigenvalue respectively.

The goal of a DP linear regression algorithm is to output a parameter θ^p that guarantees a required privacy budget and minimizes the *test risk*:

$$\mathcal{P}(\theta^p) = \frac{1}{2}\mathbb{E}_{(x,y)\sim P_{XY}}\left[\left(x^\top\theta^p - y\right)^2\right] = \frac{\left\|\Sigma^{1/2}\left(\theta^p - \theta^*\right)\right\|_2^2 + \zeta^2}{2}. \tag{9.7}$$

We also use the notation $\mathcal{R}(\theta^p) = \mathcal{P}(\theta^p) - \zeta^2/2$ to denote the excess test risk s.t. $\mathcal{R}(\theta^*) = 0$. This notation slightly differs from the simplified presentation in Section 9.1, as $\mathcal{R}$ here is a function defined in parameter space. We will stick to this notation for the rest of the chapter.

Differential privacy (DP). The definition of DP builds on the notion of *adjacent datasets*: a dataset D' is said to be adjacent to a dataset D if they differ by only one sample. In this chapter, we will frame privacy in terms of zero-concentrated DP (zCDP) [Bun and Steinke, 2016], which is defined below.

Definition 9.1 (Zero concentrated DP [Bun and Steinke, 2016])**.** *Given $\alpha \in (1, +\infty)$ and two random variables X and X' with laws p_X and $p_{X'}$, their α-Rényi Divergence [Rényi, 1961] is defined as*

$$D_\alpha(X \,\|\, X') = \frac{1}{\alpha - 1}\ln\int\left(\frac{p_X(\theta)}{p_{X'}(\theta)}\right)^\alpha p_{X'}(\theta)\mathrm{d}\theta. \tag{9.8}$$

Then, a randomized algorithm $\mathcal{A}$ satisfies $\rho^2/2$-zero concentrated DP ($\rho^2/2$-zCDP) if, for any pair of adjacent datasets D, D' and any $\alpha \in (1, +\infty)$, we have $D_\alpha(\mathcal{A}(D) \,\|\, \mathcal{A}(D')) \le \alpha\rho^2/2$.

Guarantees for zCDP can be translated to other formulations, such as (ε, δ)-DP, and we provide conversion formulas in Appendix J.1.

Algorithm 9.1: DP-GD

Input: Training data (X, Y), learning rate schedule $\{\eta_k\}_{k=1}^n$, clipping constant C_{clip}, noise multiplier schedule $\{\sigma_k\}_{k=1}^n$, initialization $\theta_0 = 0$.

for $k \in \{1, \dots, n\}$ **do**

Compute the gradient $g_k = \nabla_\theta \left(x_k^\top \theta_{k-1} - y_k\right)^2 / 2 = x_k \left(x_k^\top \theta_{k-1} - y_k\right)$; Clip the gradient $\bar{g}_k = g_k \min\left(1, \frac{C_{\text{clip}}}{\|g_k\|_2}\right)$; Set the learning rate adaptively $\bar{\eta}_k = \min\left(\eta_k, \frac{2}{\|x_k\|_2^2}\right)$; Sample independent Gaussian noise $b_k \sim \mathcal{N}(0, I)$; Update the model parameters $\theta_k = \theta_{k-1} - \bar{\eta}_k \bar{g}_k + 2C_{\text{clip}} \sigma_k b_k$;

end

Output: Model parameters $\theta^p = \theta_n$.

9.4 DP-GD and deterministic equivalent

We consider a DP gradient descent algorithm performing a single pass on the n training samples (Algorithm 9.1). In this formulation, both the learning rate η_k and the magnitude of the private noise σ_k at the k-th iteration follow the schedules $\{\eta_k\}_{k=1}^n$, $\{\sigma_k\}_{k=1}^n$ given as input, and η_k is modified adaptively as a function of $\|x_k\|_2$. The output of the algorithm is the final parameter θ_n, which is the *only* object for which we provide privacy guarantees, due to our approach based on privacy amplification by iteration [Feldman et al., 2018, 2020]. This differs from prior work where any intermediate step can be privately released [Varshney et al., 2022, Liu et al., 2023b, Brown et al., 2024b], which relies on advanced composition theorems [Dwork and Roth, 2014].

Proposition 9.2. *Algorithm 9.1 satisfies $(\rho^2/2)$-zCDP, where*

$$\rho = \max_{k \in [n]} \frac{\eta_k}{\sqrt{\sum_{j=k}^n \sigma_j^2}}. \tag{9.9}$$

Proposition 9.2 (whose proof is deferred to Appendix J.1) states that each sample x_k is "protected" by the overall noise introduced in the following updates $\sum_{j=k}^n \sigma_j^2$. For an assigned privacy guarantee, we can minimize the noise introduced by the algorithm $\sum_{j=1}^n \sigma_j^2$ (and, therefore, optimize its performance) via the schedule below:

$$\eta_k = \rho \sqrt{\sum_{j=k}^n \sigma_j^2}, \quad \text{or, equivalently,} \quad \rho^2 \sigma_k^2 = \begin{cases} \eta_k^2 - \eta_{k+1}^2, & k \in \{1, \dots, n-1\}, \\ \eta_k^2, & k = n. \end{cases} \tag{9.10}$$

Thus, from now on, we will restrict our study to the noise schedules defined in (9.10), making Algorithm 9.1 uniquely defined given the privacy parameter ρ, the clipping constant C_{clip}, and a non-increasing learning rate schedule $\{\eta_k\}_{k=1}^n$. We now make two assumptions on data distribution and hyper-parameter scaling.

Assumption 9.3 (Data distribution)**.** *$\{x_i\}_{i=1}^n$ are n i.i.d. samples from the multivariate, mean-0, Gaussian distribution $\mathcal{P}_X$, with covariance $\Sigma := \mathbb{E}\left[xx^\top\right] \in \mathbb{R}^{d \times d}$. Furthermore, the noise z_i is mean-0, Gaussian, with variance $\zeta^2 = \Theta(1) > 0$, and $\|\theta^*\|_2 = \Theta(1)$, $\lambda_{\max} = O(1)$, and $\operatorname{tr}(\Sigma) = d$.*

Gaussian data was considered in Dwork et al. [2025], while Varshney et al. [2022], Liu et al. [2023b], Brown et al. [2024b] assume bounds on the tail of the data distributions. We focus

on the Gaussian case for simplicity, but we expect to be possible to extend some of our results to data with sufficiently well-behaved tails (see Appendix A in Marshall et al. [2025]). The normalization $\mathrm{tr}(\Sigma) = d$ is also chosen to simplify the presentation of the results, and is not strictly necessary for our derivation.

Assumption 9.4 (Hyper-parameter scaling)**.** *Let the clipping constant in Algorithm 9.1 be*

$$C_{\mathsf{clip}} = c\sqrt{d}, \tag{9.11}$$

where $c > 0$ is a constant independent of n and d. Furthermore, the learning rate schedule is given by

$$\eta_k = \frac{\tilde{\eta}\left(k/n\right)}{n}, \tag{9.12}$$

where $\tilde{\eta} : [0, 1] \to \mathbb{R}$ is a function such that both $\tilde{\eta}^2\left(\cdot\right)$ and the absolute value of its first and second derivatives are uniformly bounded from above by a constant independent of n and d.

The norm of the gradient in Algorithm 9.1 is $\|g_k\|_2 = \|x_k\|_2 \left|x_k^\top \theta_{k-1} - y_k\right| = \|x_k\|_2 \sqrt{2\mathcal{P}(\theta_{k-1})}$. Then, if $\mathcal{P}(\theta_{k-1}) = \Theta(1)$, we have that $\|g_k\|_2 = \Theta(\sqrt{d})$, as $\|x_k\|_2 = \Theta(\sqrt{d})$ with high probability by Assumption 9.3. Note that the condition $C_{\mathsf{clip}} = c\sqrt{d}$ considers asymptotically smaller clipping constants than the ones considered e.g. in Brown et al. [2024b], which in this setting are of order $C_{\mathsf{clip}} = \Omega(\sqrt{d}\log^3 n)$ (see their Theorem 2.7). Let us also introduce

$$r(\theta, x, y) = x^\top \theta - y, \qquad r_c(\theta, x, y) = r(\theta, x, y) \min\left(1, \frac{c}{|r(\theta, x, y)|}\right), \tag{9.13}$$

where $r(\theta, x, y)$ represents the residual in θ, and $r_c(\theta, x, y)$ is a clipped version of it. Then, as done in Marshall et al. [2025], we define the *descent reduction factor* and the *variance reduction factor*

$$\mu_c(\theta) = \frac{\left\|\mathbb{E}_{(x,y)\sim P_{XY}}\left[r_c(\theta, x, y)\, x\right]\right\|_2}{\left\|\mathbb{E}_{(x,y)\sim P_{XY}}\left[r(\theta, x, y)\, x\right]\right\|_2}, \qquad \nu_c(\theta) = \frac{\mathbb{E}_{(x,y)\sim P_{XY}}\left[r_c(\theta, x, y)^2\right]}{\mathbb{E}_{(x,y)\sim P_{XY}}\left[r(\theta, x, y)^2\right]}. \tag{9.14}$$

In Lemmas J.4 and J.5, we show that, due to Assumption 9.3, the functions $\mu_c(\theta)$ and $\nu_c(\theta)$ only depend on the risk $\mathcal{R}(\theta)$, i.e., $\mu_c(\theta_1) = \mu_c(\theta_2)$ if $\mathcal{R}(\theta_1) = \mathcal{R}(\theta_2)$ (and same for ν_c). Then, with a slight abuse of notation, we will regard μ_c and ν_c as functions from $\mathbb{R}_{\geq 0}$ to $(0, 1)$ taking as argument directly the value of the risk. Furthermore, in the aforementioned Lemmas, we show that, for any fixed value of the risk R, $\mu_c(R)$ and $\nu_c(R)$ are monotonically increasing in c, going from 0 (as $c \to 0$) to 1 (as $c \to +\infty$), see Figure J.1.

Definition 9.5 (Deterministic equivalent)**.** *For any $t \in [0, 1)$, define the system of d coupled ODEs*

$$\mathrm{d}D_i(t) = -2\lambda_i \bar{\eta}(t)\mu_c(R(t))D_i \mathrm{d}t + \lambda_i \bar{\eta}^2(t)\nu_c(R(t))(R(t) + \zeta^2/2)\gamma_n \mathrm{d}t + 2c^2\tilde{\sigma}^2(t)\gamma_n^2 \mathrm{d}t, \tag{9.15}$$

with $\gamma_n = d/n$, $\bar{\eta}(t) = \min(\tilde{\eta}(t), 2/\gamma_n)$, $\tilde{\sigma}(t)$ such that $\rho^2\tilde{\sigma}^2(t) = -\mathrm{d}\tilde{\eta}^2(t)/\mathrm{d}t$,

$$R(t) = \frac{1}{d}\sum_{i=1}^{d} \lambda_i D_i(t), \tag{9.16}$$

and initial condition $D_i(0) = d\left(\omega_i^\top \theta^*\right)^2/2$.

We remark that the previous system of ODEs has a unique solution, due to Picard–Lindelöf theorem (see Chapter 2 in Teschl [2012]), since the RHS of (9.15) is continuous in t, locally Lipschitz in D_i ($\mu_c(R)$ and $\nu_c(R)$ are bounded and Lipschitz due to Lemma J.4), and since it respects the linear growth condition in Theorem 2.17 in Teschl [2012].

Theorem 9.6. *Let Assumptions 9.3 and 9.4 hold. Let $\rho = \Theta(1)$, and $n, d \to \infty$ s.t. $\gamma_n = d/n \to \gamma \in (0, \infty)$. Denote by θ_k a realization of Algorithm 9.1, and by $R(t)$ the solution to the system of ODEs described by* (9.15) *and* (9.16)*. Then, with overwhelming probability, we have*

$$\sup_{t\in[0,1)} \left|\mathcal{R}(\theta_{\lfloor tn\rfloor}) - R(t)\right| = O\left(\frac{\log^2 n}{\sqrt{n}}\right). \tag{9.17}$$

Furthermore, we also have

$$\sup_{t\in[0,1)} \left|\mathbb{E}\left[\mathcal{R}(\theta_{\lfloor tn\rfloor})\right] - R(t)\right| = O\left(\frac{\log^2 n}{\sqrt{n}}\right), \tag{9.18}$$

where the expectation is taken with respect to both the data and the algorithm.

In words, Theorem 9.6 states that the risk of Algorithm 9.1 can be well approximated in terms of the solution of a system of d coupled deterministic ODEs. This system of ODEs then provides a deterministic equivalent for the dynamics of DP-GD, since it approximates sharply its risk without depending on the stochasticity of the algorithm and of the data. Definition 9.5 and Theorem 9.6 also implicitly introduce the continuous time variable $t \in [0, 1]$, which is mapped to the iterations via $k = \lfloor tn \rfloor$.

Notice that, in the isotropic case where $\Sigma = I$, the system in (9.15) reduces to the single ODE

$$\mathsf{d}R(t) = -2\bar{\eta}(t)\mu_c(R(t))R(t)\mathsf{d}t + \bar{\eta}^2(t)\nu_c(R(t))(R(t) + \zeta^2/2)\gamma_n\mathsf{d}t + 2c^2\tilde{\sigma}^2(t)\gamma_n^2\mathsf{d}t, \tag{9.19}$$

with $R(0) = \|\theta^*\|_2^2/2$, $\bar{\eta}(t) = \min(\tilde{\eta}(t), 2/\gamma_n)$, and $\rho^2\tilde{\sigma}^2(t) = -\mathsf{d}\tilde{\eta}^2(t)/\mathsf{d}t$. The RHS of (9.19) has three terms: *(i)* a negative term proportional to $R(t)$, which captures the descent towards the minimizer of R, *(ii)* a positive term proportional to $R(t) + \zeta^2/2$, which captures the fact that Algorithm 9.1 at step k does not optimize R directly, but rather an estimate based on (x_k, y_k), and *(iii)* another positive term proportional to $\tilde{\sigma}^2(t)$, which models the private noise in Algorithm 9.1. We also remark that the first two terms on the RHS of (9.19) are proportional to $\bar{\eta}(t) = \min(\tilde{\eta}(t), 2/\gamma_n)$, in analogy to the adaptive learning rate step that defines $\bar{\eta}_k$ in Algorithm 9.1.

For general covariance, it is possible to upper and lower bound the value of $R(t)$ as a function of the largest ($\lambda_{\max}$) and smallest ($\lambda_{\min}$) eigenvalues of the covariance matrix Σ. In particular, defining

$$\begin{aligned}
\mathsf{d}\overline{R}(t) &= -2\lambda_{\min}\tilde{\eta}(t)\mu_c(\overline{R})\overline{R}\mathsf{d}t + \lambda_{\max}\tilde{\eta}^2(t)\nu_c(\overline{R})(\overline{R} + \zeta^2/2)\gamma_n\mathsf{d}t + 2c^2\tilde{\sigma}^2(t)\gamma_n^2\mathsf{d}t,\\
\mathsf{d}\underline{R}(t) &= -2\lambda_{\max}\tilde{\eta}(t)\mu_c(\underline{R})\underline{R}\mathsf{d}t + \tilde{\eta}^2(t)\nu_c(\underline{R})(\underline{R} + \zeta^2/2)\gamma_n\mathsf{d}t + 2c^2\tilde{\sigma}^2(t)\gamma_n^2\mathsf{d}t,
\end{aligned} \tag{9.20}$$

with $\overline{R}(0) = \underline{R}(0) = R(0) = \|\Sigma^{1/2}\theta^*\|_2^2/2$, we have, for every $t \in [0, 1]$,

$$\underline{R}(t) \le R(t) \le \overline{R}(t). \tag{9.21}$$

We refer to Proposition J.12 in Appendix J.4 for the formal statement and proof. Compared to the isotropic case, the upper bound $\overline{R}(t)$ reduces the descent term by a factor $\lambda_{\min} \le 1$

and increases the second term in the RHS of (9.19) by a factor $\lambda_{\max} \geq 1$. Instead, the lower bound $\underline{R}(t)$ just increases the descent term by a factor $\lambda_{\max} \geq 1$. In Section 9.5 we use these bounds to *(i)* quantify the benefit of clipping, *(ii)* establish the minimax optimality of DP-GD with a properly chosen learning rate schedule and, more generally, to *(iii)* give convergence rates for a wide class of schedules. In Section 9.6 we consider directly the system of d ODEs in (9.15) to *(iv)* derive scaling laws.

Remark 9.7. The risk of Algorithm 9.1 can be compactly characterized also through the solution of the following stochastic differential equation

$$\mathrm{d}\Theta_t = -\bar{\eta}(t)\mu_c(\Theta_t)\nabla\mathcal{P}(\Theta_t)\mathrm{d}t + \bar{\eta}(t)\sqrt{\frac{2\nu_c(\Theta_t)\mathcal{P}(\Theta_t)\Sigma}{n}}\mathrm{d}B_t^s + 2\frac{\sqrt{d}}{n}c\tilde{\sigma}(t)\mathrm{d}B_t^p, \tag{9.22}$$

where $\Theta_0 = \theta_0 = 0$, B_t^s and B_t^p are two independent standard Brownian motions in $\mathbb{R}^d$, $\bar{\eta}(t) = \min(\tilde{\eta}(t), 2n/d)$, and $\tilde{\sigma}(t)$ is such that $\rho^2\tilde{\sigma}^2(t) = -\mathrm{d}\tilde{\eta}^2(t)/\mathrm{d}t$. More precisely, in Appendix J.3, we show that, with overwhelming probability,

$$\sup_{t\in[0,1)} \left|\mathcal{R}(\theta_{\lfloor tn\rfloor}) - \mathcal{R}(\Theta_t)\right| = O\left(\frac{\log^2 n}{\sqrt{n}}\right).$$

This is a description of the dynamics of DP-GD through a single stochastic differential equation, rather than through a system of d coupled ODEs.

Proof ideas for Theorem 9.6. The argument follows a similar strategy as the one proving Theorem 1 in Marshall et al. [2025]. In particular, let $u_k = \theta_k - \theta^*$ and $V_t = \Theta_t - \theta^*$, where Θ_t is defined in (9.22). Define the set of quadratic functions

$$q_z(v) = \frac{1}{2}v^\top(\Sigma - zI)^{-1}v, \qquad z \in \Omega = \{z \in \mathbb{C} : |z| = 2\left\|\Sigma\right\|_{\mathsf{op}}\}. \tag{9.23}$$

On the one hand, we can write the Doob's decomposition of the update

$$q_z(u_{k+1}) - q_z(u_k) = \frac{1}{n}\mathcal{F}_z\left(u_k, R(u_k), k/n\right) + \Delta M_k(z) + \Delta E_k(z). \tag{9.24}$$

Here, $\mathcal{F}_z$ denotes the predictable drift, with contributions coming from the deterministic descent term, the sampling-variance term, and the private-noise term (see (J.40)):

$$\begin{aligned}\mathcal{F}_z(u, R, t) = &-\bar{\eta}(t)\mu_c(R)\left(\|u\|_2^2 + zq_z(u)\right) + \bar{\eta}(t)^2\nu_c(R)\left(R + \zeta^2/2\right)\gamma_n\frac{\mathrm{tr}\left(\Sigma(\Sigma - zI)^{-1}\right)}{d} \\ &+ 2c^2\tilde{\sigma}^2(t)\gamma_n^2\frac{\mathrm{tr}\left((\Sigma - zI)^{-1}\right)}{d}.\end{aligned} \tag{9.25}$$

Combining Lemma 2 in Marshall et al. [2025] with Lemmas J.8 and J.11, we show that the martingale and error terms in (9.24) ($\sum_{j\leq k}\Delta M_j(z)$ and $\sum_{j\leq k}\Delta E_j(z)$) are $\tilde{O}(n^{-1/2})$.

On the other hand, applying Itô's formula to (9.22) yields

$$\mathrm{d}q_z(V_t) = \mathcal{F}_z\left(V_t, R(\Theta_t), t\right)\mathrm{d}t + \mathrm{d}M_t^{\mathrm{sde}}(z). \tag{9.26}$$

Here, the predictable part $\mathcal{F}_z$ matches with the one of the discrete dynamics and, in Lemma J.9, we show that the remaining term is small, i.e., $|\int_0^t \mathrm{d}M_s^{\mathrm{sde}}(z)| = \tilde{O}(n^{-1/2})$ with overwhelming probability. Then, relying on the Lipschitz continuity of $\mathcal{F}_z$ with respect to its first argument

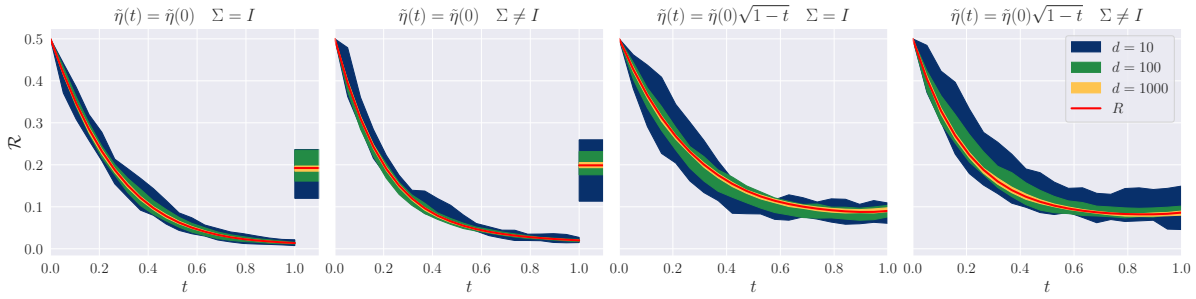


Figure 9.1: Numerical simulations for Algorithm 9.1 ($d \in \{10, 100, 1000\}$) and the solution $R(t)$ of the system of ODEs in (9.15) and (9.16). We consider the two schedules $\tilde\eta(t) = \tilde\eta(0)$ (first and second panel) and $\tilde\eta(t) = \tilde\eta(0)\sqrt{1-t}$ (third and fourth panel). We fix $\gamma = 0.1$, $\rho = 1$, $\zeta = 0.3$, $(\theta_i^*)^2 = 1/d$, $\tilde\eta(0) = 3$, $c = 1$, and consider both isotropic data ($\Sigma = I$, first and third panel) and data with diagonal covariance whose eigenvalues are uniformly distributed in the interval $[0, 2]$ (second and fourth panel). For $\alpha = 0$, we also report for $t \geq 1$ the risk $\mathcal{R}(\theta^p)$, and the value of $R(1) + 2c^2\tilde\eta^2(1)\gamma^2/\rho^2$ with a red line. For each value of d, we report bands corresponding to 1 standard deviation around the mean over 10 independent trials of Algorithm 9.1.

and controlling the discretization errors, Gronwall's inequality yields (see the argument in (J.45)-(J.48))

$$\sup_{t\in[0,1)} \left| q_z(u_{\lfloor tn \rfloor}) - q_z(V_t) \right| = \tilde{O}\left(n^{-1/2}\right), \tag{9.27}$$

with overwhelming probability uniformly in $z \in \Omega$, due to a net argument on Ω. Now, intuitively, one can remove the martingale term from (9.26) and consider the resulting deterministic evolution. As

$$q_z(V) = \frac{1}{2}\sum_{i=1}^{d} \frac{(V^\top \omega_i)^2}{\lambda_i - z}, \qquad \|V\|_2^2 = \frac{i}{\pi}\oint_\Omega q_z(V)\mathrm{d}z, \qquad R(V+\theta^*) = \frac{i}{2\pi}\oint_\Omega z q_z(V)\mathrm{d}z, \tag{9.28}$$

where the last two equalitites follow from the Cauchy integral formula, we have that $q_z(V_t)$ is described by an integro-differential equation in z and t (see (J.78)). Then, due to Lemma 4.1 in Collins-Woodfin et al. [2024a], its unique solution is

$$q_z(V_t) = \frac{1}{d}\sum_{i=1}^{d} \frac{D_i(t)}{\lambda_i - z}, \tag{9.29}$$

where $D_i(t)$ is defined in (9.15). Using the last relation in (9.28) recovers $R(t) = \sum_{i=1}^d \lambda_i D_i(t)/d$, which gives the desired result. The full argument is deferred to Appendix J.3.

Noise in the last iteration. Theorem 9.6 holds for $\{\theta_k\}_{k=1}^{n-1}$, as the supremum in (9.17) is taken on the open interval $t \in [0, 1)$, and in general the equivalence *does not hold* for the last iterate θ_n (corresponding to $t = 1$), which gives the (private) output of the algorithm. Intuitively, for $k < n$, (9.10) suggests (with abuse of notation)

$$\rho^2\sigma_k^2 = -\frac{\mathsf{d}}{\mathrm{d}k}\eta_k^2 = -\frac{1}{n}\frac{\mathsf{d}}{\mathsf{d}\,(k/n)}\frac{\tilde\eta^2\,(k/n)}{n^2} \approx \frac{\rho^2}{n^3}\tilde\sigma^2(k/n), \tag{9.30}$$

where in the second step we use Assumption 9.4 and $\tilde\sigma$ is given by Definition 9.5. In contrast, for the last iterate, again due to (9.10), we have $\rho^2\sigma_n^2 = \eta_n^2 = \tilde\eta^2(1)/n^2$. The additional n factor in the denominator of (9.30) implies that, if $\tilde\eta(1) > 0$, the noise added to the last iterate is much larger. Thus, we treat the case separately via the result below (proved in Appendix J.3), which quantifies the final test loss of $\theta_n = \theta^p$.

Proposition 9.8. *Let Assumptions 9.3 and 9.4 hold. Let* $\rho = \Theta(1)$ *and* $n, d \to \infty$ *s.t.* $\gamma_n = d/n \to \gamma \in (0, \infty)$. *Then, with overwhelming probability, we have*

$$\left| \mathcal{R}(\theta^p) - R(1) - \frac{2c^2\tilde{\eta}^2(1)\gamma_n^2}{\rho^2} \right| = O\left(\frac{\log^2 n}{\sqrt{n}}\right). \tag{9.31}$$

Furthermore, we also have

$$\left| \mathbb{E}\left[\mathcal{R}(\theta^p)\right] - R(1) - \frac{2c^2\tilde{\eta}^2(1)\gamma_n^2}{\rho^2} \right| = O\left(\frac{\log^2 n}{\sqrt{n}}\right). \tag{9.32}$$

The convergence predicted by Theorem 9.6 and Proposition 9.8 is already evident at moderate values of n, d, as showcased by Figure 9.1 for different schedules ($\tilde{\eta}(t) = \tilde{\eta}(0)$ and $\tilde{\eta}(t) = \tilde{\eta}(0)\sqrt{1-t}$) and different data covariances. In the panels for $\tilde{\eta}(t) = \tilde{\eta}(0)$ of Figure 9.1, we report in a short interval at $t \geq 1$ the risk $\mathcal{R}(\theta^p)$, after the noise in the last iteration is added, since $\tilde{\eta}(1) > 0$.

9.5 Optimal rates

Putting together Theorem 9.6 and Proposition 9.8 gives that the final risk $\mathcal{R}(\theta^p)$ asymptotically converges to a constant independent of n, d and only dependent on γ. This is in sharp contrast with earlier upper bounds by Varshney et al. [2022], Liu et al. [2023b], Brown et al. [2024b] which diverge in the proportional regime (n scaling linearly in d) considered here. Notably, this is a consequence of taking $C_{\mathsf{clip}} = c\sqrt{d}$ (see Assumption 9.4) with c a constant independent of n, d, rather than $C_{\mathsf{clip}} = \omega(\sqrt{d})$.

We now show that the characterization put forward in Section 9.4 is precise enough to capture how $\mathcal{R}(\theta^p)$ depends on γ and ρ, as well as on the algorithm hyper-parameters $\tilde{\eta}(t)$ and c. While Dwork et al. [2025] also derive a result which does not diverge in the proportional regime, their analysis does not draw statistical implications on how the final risk depends on sample size and privacy requirement. One significant challenge towards this goal is that, in the framework of Dwork et al. [2025], the trajectory of DP-GD is expressed via complex DMFT equations, which are then hard to analyze precisely.

Throughout the section, we focus on the non-asymptotic behavior of $\mathcal{R}(\theta^p)$ as $\gamma \to 0$. Importantly, the limit $\gamma \to 0$ is taken *after* the limit $d, n \to \infty$, which informally means that d and n are incomparably larger than $1/\gamma$. Thus, our bounds on $\mathcal{R}(\theta^p)$ neglect the smaller terms (vanishing as $d, n \to \infty$) from Theorem 9.6 and Proposition 9.8, including the difference $|\gamma - \gamma_n| = o(1)$. We assume that $\lambda_{\min}, \lambda_{\max}, \|\theta^*\|_2, \zeta = \Theta_\gamma(1)$, i.e., the condition number of the covariance, the test risk of the model at initialization and the label noise variance are strictly positive constants independent of γ; instead, the privacy parameter ρ is allowed to depend on γ (but not on n, d).

9.5.1 Output perturbation and constant private noise

In this section, we consider the family of learning rate schedules

$$\tilde{\eta}(t) = \tilde{\eta}(0)(1-t)^\alpha, \tag{9.33}$$

for $\alpha = 0$, $\alpha = 1/2$ and $\alpha \geq 1$. Taking $\alpha = 0$ corresponds to output perturbation: the learning rate is fixed during the whole training, and the private noise is added only at the end

of the algorithm, as (9.10) implies $\sigma_k = 0$ for $k \in \{1, \dots, n-1\}$ and $\sigma_n = \eta_1/\rho = \tilde{\eta}(0)/(n\rho)$. Taking $\alpha = 1/2$ corresponds to a linearly decaying $\tilde{\eta}^2(t)$, which in turn implies a constant level of noise $\sigma_k = \eta_1/(\sqrt{n}\rho) = \tilde{\eta}(0)/(n^{3/2}\rho)$ in the iterations of DP-GD. These two choices of learning rate schedule are analyzed in Proposition 9.9. At the end of the section, we also study the effect of taking $\alpha \geq 1$ in Proposition 9.10, so that the private noise decays with time. Note that all these values of α are in agreement with Assumption 9.4.

Proposition 9.9. *Let Assumptions 9.3 and 9.4 hold. Let θ_0^p and $\theta_{1/2}^p$ be the solutions obtained with Algorithm 9.1 with $\tilde{\eta}(t)$ given by (9.33) for $\alpha = 0$ and $\alpha = 1/2$ respectively, with $\tilde{\eta}(0) \leq 2/\gamma$. Consider $\lambda_{\min}, \lambda_{\max}, \|\theta^*\|_2, \zeta = \Theta_\gamma(1)$, and $\rho = \Omega_\gamma\left(\gamma^{1-h}\right)$ for some $h > 0$. Pick*

$$c = O_\gamma(1), \qquad \tilde{\eta}(0)c = C\ln(1/\gamma), \tag{9.34}$$

for a large enough constant C which does not depend on γ. Then, we have that, with overwhelming probability,

$$\begin{aligned} \mathcal{R}(\theta_0^p) &= O_\gamma\left(\gamma\ln(1/\gamma) + \frac{\gamma^2\ln^2(1/\gamma)}{\rho^2}\right), \\ \mathcal{R}(\theta_{1/2}^p) &= O_\gamma\left(\gamma\ln^{2/3}(1/\gamma) + \frac{\gamma^2\ln^{4/3}(1/\gamma)}{\rho^2}\right). \end{aligned} \tag{9.35}$$

Furthermore, for any choice of the hyper-parameters c and $\tilde{\eta}(0)$, we have that

$$\begin{aligned} \mathcal{R}(\theta_0^p) &= \Omega_\gamma\left(\gamma\ln(1/\gamma) + \frac{\gamma^2\ln^2(1/\gamma)}{\rho^2}\right), \\ \mathcal{R}(\theta_{1/2}^p) &= \Omega_\gamma\left(\gamma\ln^{2/3}(1/\gamma) + \frac{\gamma^2\ln^{4/3}(1/\gamma)}{\rho^2}\right). \end{aligned} \tag{9.36}$$

Proposition 9.9 (proved in Appendix J.4, also in the setting where $\lambda_{\max}$ and $\lambda_{\min}$ can depend on γ) gives upper and lower bounds for output perturbation ($\alpha = 0$) and DP-GD with constant noise ($\alpha = 1/2$). This result has two remarkable consequences: *(i)* the hyper-parameters in (9.34) are optimal in terms of rate, and *(ii)* DP-GD with constant noise outperforms output perturbation when γ is sufficiently small. After giving a proof sketch, we elaborate on these two points in the remainder of the section.

Proof ideas for Proposition 9.9. For simplicity, we consider here the isotropic case (the case with general covariance is treated in the full argument in Appendix J.4). The ODE in (9.19) reads

$$\mathrm{d}R(t) = -2\tilde{\eta}(0)c\frac{\mu_c(R(t))}{c}R(t)\mathrm{d}t + (\tilde{\eta}(0)c)^2\frac{\nu_c(R(t))(R(t)+\zeta^2/2)}{c^2}\gamma\mathrm{d}t. \tag{9.37}$$

If $c = O(1)$, the bounds on $\mu_c(R)$ and $\nu_c(R)$ in Lemma J.5 yield that, uniformly in $t \in [0,1]$,

$$\frac{\mu_c(R(t))}{c} = \Theta_\gamma(1), \qquad \frac{\nu_c(R(t))(R(t)+\zeta^2/2)}{c^2} = \Theta_\gamma(1). \tag{9.38}$$

Thus, due to ODE comparison arguments, we can characterize $R(t)$ studying the closed-form solution of the ODE

$$\mathrm{d}\tilde{R}(t) = -\tilde{\eta}(0)c\tilde{R}(t)\mathrm{d}t + (\tilde{\eta}(0)c)^2\gamma\mathrm{d}t, \tag{9.39}$$

with $\tilde{R}(0) = R(0)$, i.e., $\tilde{R}(t) = (R(0) - \gamma\tilde{\eta}(0)c)\, e^{-\tilde{\eta}(0)ct} + \gamma\tilde{\eta}(0)c$. Then, setting $t = 1$, Proposition 9.8 gives that the risk is well approximated by

$$(R(0) - \gamma\tilde{\eta}(0)c)\, e^{-\tilde{\eta}(0)c} + \gamma\tilde{\eta}(0)c + \frac{(\tilde{\eta}(0)c)^2\gamma^2}{\rho^2}. \tag{9.40}$$

Minimizing over $\tilde{\eta}(0)c$ gives the desired result.

If $c = \Omega(1)$, the bounds in Lemma J.5 yield

$$\mu_c(R(t)) = \Theta_\gamma(1), \qquad \nu_c(R(t))(R(t) + \zeta^2/2) = \Omega_\gamma(1). \tag{9.41}$$

Then, $R(t)$ is lower bounded (up to constants independent of γ) by the solution to the ODE

$$\mathsf{d}\tilde{R}(t) = -\tilde{\eta}(0)\tilde{R}(t)\mathsf{d}t + \tilde{\eta}(0)^2\gamma\mathsf{d}t. \tag{9.42}$$

This ODE has the same form as in (9.39) (after replacing $\tilde{\eta}(0)c$ with just $\tilde{\eta}(0)$) and, therefore, it does not give a lower final risk.

The case $\alpha = 1/2$ follows a similar strategy. In particular, if $c = O(1)$, $R(t)$ can be characterized via the ODE

$$\mathsf{d}\tilde{R}(t) = -\tilde{\eta}(0)c\sqrt{1-t}\tilde{R}(t)\mathsf{d}t + (\tilde{\eta}(0)c)^2\gamma(1-t)\mathsf{d}t + \frac{(\tilde{\eta}(0)c)^2\gamma^2}{\rho^2}\mathsf{d}t, \tag{9.43}$$

with $\tilde{R}(0) = R(0)$. The solution to (9.43) can also be written in closed-form (see (J.152)). Thus, setting $\tilde{\eta}(0)c$ according to (9.34) allows to obtain the improved upper bound.

Optimal hyper-parameters. We now comment on the optimal hyper-parameter choice in (9.34). First, notice that if $\tilde{\eta}(0) > 2/\gamma$, the gradient update would be proportional to $\bar{\eta}_k < \eta_k$, while the private noise σ_k still depends on η_k via (9.10). This suggests the sub-optimality of having $\bar{\eta}_k < \eta_k$ and of the regime $\tilde{\eta}(0) > 2/\gamma$. Second, the choice $\tilde{\eta}(0)c = C\ln(1/\gamma)$ provides the optimal trade-off between two competing objectives: on the one hand, $\tilde{\eta}(t)c$ controls the size of the first term of the ODEs in (9.20) (for $c = O_\gamma(1)$ and bounded values of R, Lemma J.5 gives that $\mu_c(R)/c$ is lower bounded by a constant), which in turn determines the speed of convergence towards 0 of the risk; on the other hand, a large product $\tilde{\eta}(0)c$ increases the last two terms of the ODEs, which have the opposite effect of increasing the risk. Third, the choice $c = O_\gamma(1)$ is motivated by the fact that increasing c beyond this point does not further increase $\mu_c(R) < 1$, which drives the risk to 0. However, larger values of c increase the private noise in DP-GD, and hence the last term in the RHSs of (9.20), which increases the risk.

These conclusions are supported by Figures 9.4, 9.5 and 9.6: performance deteriorates if either c exceeds 1 (upper part of the heatmaps) or $\tilde{\eta}(0)$ exceeds $2/\gamma$ (right part of the heatmaps); the lowest values of the risk are roughly parallel to the line $c\tilde{\eta}(0) = \ln(1/\gamma)$. A more detailed discussion of the numerical experiments is in Section 9.7.

Benefits of aggressive clipping. Adding the further condition $\rho/c = \Omega(\gamma^{1-h})$ for some $h > 0$, the lower bounds in Proposition 9.9 can be improved to

$$\mathcal{R}(\theta^p_{0,1/2}) = \tilde{\Omega}_\gamma\left(\gamma + \frac{\max(1, c^2)\gamma^2}{\rho^2}\right), \tag{9.44}$$

where $\tilde{\Omega}_\gamma$ neglects poly-logarithmic factors. The term $\max(1, c^2)$ in the second quantity at the RHS of (9.44) quantifies the negative effects on performance due to using a large clipping

constant $c = \omega_\gamma(1)$. This is also evident from the numerical results in Figures 9.4, 9.5, and 9.6. We note that the requirement $\rho/c = \Omega(\gamma^{1-h})$ is analogous to the lower bound for ρ in the statement of Proposition 9.9, and it guarantees that the second term in the RHS of (9.44) is $o_\gamma(1)$, making it comparable to the first term in the RHS.

The lower bound in (9.44) is formally discussed at the end of the proof of Proposition 9.9 in Appendix J.4.1, and it informally follows from the argument that leads to (9.42), which shows that the first two terms in the RHS of the ODE in (9.19) are qualitatively identical in the regimes $c = O_\gamma(1)$ and $c = \Omega_\gamma(1)$. In the first case ($c = O_\gamma(1)$), these two terms are proportional to $\tilde{\eta}(t)c$ and $(\tilde{\eta}(t)c)^2$; in the second case ($c = \Omega_\gamma(1)$), they are proportional to $\tilde{\eta}(t)$ and $\tilde{\eta}(t)^2$. Thus, with a change of variable, (9.19) behaves as the following ODE:

$$\mathsf{d}\tilde{R}(t) = -v(t)\tilde{R}(t)\mathsf{d}t + v^2(t)\gamma\mathsf{d}t + \frac{\max(1, c^2)}{\rho^2}\left(-\frac{\mathsf{d}v^2(t)}{\mathsf{d}t}\right)\gamma^2\mathsf{d}t, \tag{9.45}$$

where $v(t) = \tilde{\eta}(t)c$ in the case $c = O_\gamma(1)$, and $v(t) = \tilde{\eta}(t)$ in the case $c = \Omega_\gamma(1)$.

Benefits of decaying the learning rate. Motivated by the improvement of DP-GD with constant noise over output perturbation (Proposition 9.9), we consider reducing the noise added at the end of training, i.e., we pick a polynomial schedule as in (9.33) with $\alpha \geq 1$. This implies that $\tilde{\sigma}^2(t)$ is proportional to $2\alpha(1-t)^{2\alpha-1}$, which corresponds to decaying the noise at least linearly.

Proposition 9.10. *Let Assumptions 9.3 and 9.4 hold. Let θ^p_α be the solution obtained with Algorithm 9.1, with $\tilde{\eta}(t)$ given by* (9.33) *for $\alpha \geq 1$, with $\tilde{\eta}(0) \leq 2/\gamma$. Consider $\lambda_{\min}, \lambda_{\max}, \|\theta^*\|_2, \zeta = \Theta_\gamma(1)$, and*

$$\ln^2(1/\gamma)\gamma\left(\alpha + \frac{\gamma}{\rho^2}\right) = o_\gamma(1). \tag{9.46}$$

Pick

$$c = O_\gamma(1), \qquad \tilde{\eta}(0)c = C\alpha\ln(1/\gamma), \tag{9.47}$$

for a large enough constant C which does not depend on γ. Then, we have that, with overwhelming probability,

$$\mathcal{R}(\theta^p_\alpha) = O_\gamma\left(\alpha\gamma\ln^{\frac{1}{1+\alpha}}(1/\gamma) + \frac{\alpha^2\gamma^2\ln^{\frac{2}{1+\alpha}}(1/\gamma)}{\rho^2}\right). \tag{9.48}$$

Proposition 9.10 (proved in Appendix J.4.2, also in the setting where $\lambda_{\max}$ and $\lambda_{\min}$ can depend on γ) provides upper bounds on the final risk as the degree α of the polynomial schedule changes, matching the ones in Proposition 9.9 for $\alpha = \{0, 1/2\}$. We note that the condition in (9.46) allows for values of α as large as $1/\gamma$ (neglecting poly-logarithmic factors), and it imposes a similar lower bound for ρ as the one in Proposition 9.9.

An immediate consequence of Proposition 9.10 is that, by taking $\alpha = \ln\ln(1/\gamma)$, one has

$$\mathcal{R}(\theta^p) = O_\gamma\left(\gamma\ln\ln(1/\gamma) + \frac{\gamma^2}{\rho^2}(\ln\ln(1/\gamma))^2\right). \tag{9.49}$$

Thus, for sufficiently small γ, it is convenient to increase α and decay the noise faster during training, up to a level $\alpha = \Theta_\gamma(\ln\ln(1/\gamma))$. Values of α larger than that may then deteriorate performance. This is illustrated in Figure 9.7 (discussed in detail in Section 9.7), which compares different schedules after the hyper-parameters $\tilde{\eta}(0)$ and c have been optimized numerically.

9.5.2 Harmonic schedule and minimax rates

Next, we consider a harmonically decreasing schedule of the form $\tilde{\eta}(t) = \beta/(t+\tau)$.

Theorem 9.11. *Let Assumptions 9.3 and 9.4 hold. Consider $\lambda_{\min}, \lambda_{\max}, \|\theta^*\|_2, \zeta = \Theta_\gamma(1)$, and $\gamma^2/\rho^2 = o_\gamma(1)$. Let θ_h^p be the solution of Algorithm 9.1 with learning rate schedule*

$$\tilde{\eta}(t) = \frac{\beta}{t+\tau}. \tag{9.50}$$

Then, there exist values of c, τ and β such that, with overwhelming probability,

$$\mathcal{R}(\theta_h^p) = O_\gamma\left(\gamma + \frac{\gamma^2}{\rho^2}\right), \tag{9.51}$$

with the same upper-bound also holding for $\mathbb{E}\left[\mathcal{R}(\theta^p)\right]$.

This result shows that the harmonically decreasing schedule $\tilde{\eta}(t) = \beta/(t+\tau)$, for an appropriate choice of β and τ, allows to get rid of the $\ln\ln(1/\gamma)$ factors in (9.49), guaranteeing the upper bound in (9.51). More precisely, β and c are chosen as constants independent of γ, but dependent on $\lambda_{\min}, \lambda_{\max}, \|\theta^*\|_2, \zeta$, while τ scales with the maximum between γ and γ^2/ρ^2. The exact values are provided in the proof of Theorem 9.11 in Appendix J.4.3, which also covers the setting where $\lambda_{\max}$ and $\lambda_{\min}$ can depend on γ.

Proof ideas for Theorem 9.11. To give an intuition of the benefits of a harmonically decreasing schedule, let us consider the simplified ODE

$$\mathsf{d}\tilde{R}(t) = -\tilde{\eta}(t)c\tilde{R}(t)\mathsf{d}t + (\tilde{\eta}(t)c)^2\gamma\mathsf{d}t. \tag{9.52}$$

We note the resemblance between (9.52) and (9.39), which we used to study the constant learning rate schedule. Heuristically, to minimize $\tilde{R}(1)$, we wish to minimize the RHS point-wise in t. Setting to 0 the derivative of the RHS with respect to $\tilde{\eta}(t)c$, we get

$$\tilde{R}(t) = 2\tilde{\eta}(t)c\gamma. \tag{9.53}$$

Plugging this into (9.52), we obtain

$$\mathsf{d}\tilde{R}(t) = -\frac{\tilde{R}^2(t)}{4\gamma}\mathsf{d}t \quad \Rightarrow \quad R(t) = \frac{4\gamma}{t + 4\gamma/\tilde{R}(0)}, \tag{9.54}$$

which together with (9.53) gives the harmonic schedule $\tilde{\eta}(t)c = \frac{2}{t+4\gamma/\tilde{R}(0)}$.

Consider now the original ODE in (9.19). Via the same comparison arguments used to obtain (9.43), when $c = O_\gamma(1)$ this ODE can be reduced to

$$\mathsf{d}\hat{R}(t) = -\tilde{\eta}(t)c\hat{R}(t)\mathsf{d}t + (\tilde{\eta}(t)c)^2\gamma\mathsf{d}t - \frac{\mathsf{d}(\tilde{\eta}(t)c)^2}{\mathsf{d}t}\frac{\gamma^2}{\rho^2}\mathsf{d}t, \tag{9.55}$$

with initial condition $\hat{R}(0) = R(0)$. Then, we have that (see Lemma J.16)

$$\hat{R}(t) = e^{-F(t)}\left(R(0) + \int_0^t e^{F(s)}\left(\gamma(\tilde{\eta}(s)c)^2 - 2\frac{\gamma^2}{\rho^2}\tilde{\eta}(s)\tilde{\eta}'(s)c^2\right)\mathsf{d}s\right), \tag{9.56}$$

where $\tilde{\eta}'(t)$ denotes the derivative of $\tilde{\eta}(t)$ and $F(t) = \int_0^t \tilde{\eta}(s)c\,\mathsf{d}s$. Relying on the previous harmonic heuristic and carrying out explicit computations after plugging the schedule in (9.50) yields the desired result.

Minimax rates. Given the result of Theorem 9.11, it is natural to question if a different choice of schedule, or even a completely different choice of algorithm could improve the rates in (9.51). In the following, we prove that this is not the case.

Theorem 9.12. *Let Assumption 9.3 hold. Consider* $\lambda_{\min}, \lambda_{\max}, \zeta^2 = \Theta_\gamma(1)$ *and* $\gamma^2/\rho^2 = O_\gamma(1)$. *Let* θ^p *be the solution found by a generic algorithm* $\mathcal{M}$ *belonging to the set of all* $\rho^2/2$*-zCDP algorithms* $\mathbb{M}$. *Then, we have*

$$\inf_{\mathcal{M}\in\mathbb{M}} \sup_{\|\theta^*\|_2<1} \mathbb{E}\left[\mathcal{R}(\theta^p)\right] = \Omega_\gamma\left(\gamma + \frac{\gamma^2}{\rho^2}\right). \tag{9.57}$$

The proof of Theorem 9.12 follows the same tracing attack technique considered in Cai et al. [2021]. We note that Theorem 4.1 in Cai et al. [2021] is stated for (ε, δ) approximate DP and, if applied directly to our setting, would yield an extra $\log n$ at the denominator of the privacy term. To avoid this, our minimax rates concern the set of $\rho^2/2$-zCDP algorithms, as per Definition 9.1. This gives a minimax lower bound matching the upper bound of Theorem 9.11.

Proof ideas for Theorem 9.12. We define the random variable $L_i = p_{\mathcal{M}(D)}(\theta)/p_{\mathcal{M}(D_i')}(\theta)$, where $p_{\mathcal{M}(D)}$ and $p_{\mathcal{M}(D_i')}$ are the probability density functions of the parameters $\mathcal{M}(D)$ and $\mathcal{M}(D_i')$, with D and D_i being neighboring datasets differing in their i-th entry. Due to the definition of zCDP (see Definition 9.1), we have that, for all $\alpha > 1$,

$$\mathbb{E}\left[L_i^\alpha\right] \leq \exp\left(\frac{\alpha(\alpha-1)\rho^2}{2}\right). \tag{9.58}$$

Consider now the tracing attack

$$A_i(\theta) = z_i x_i^\top (\theta - \theta^*). \tag{9.59}$$

Then, in Lemma J.17 (similar to Lemma 4.1 in Cai et al. [2021]) we show that there exists a prior π on θ^* with support in $\|\theta^*\|_2 < 1$ such that, for any $\rho^2/2$-zCDP algorithm $\mathcal{M}$,

$$\mathbb{E}_\pi\left[\sum_{i\in[n]} \mathbb{E}\left[A_i(\mathcal{M}(D))\right]\right] = \Omega(d). \tag{9.60}$$

In Lemma J.18 we show that this bound, together with (9.58) for $\alpha = 2$, gives

$$\inf_{\mathcal{M}\in\mathbb{M}} \mathbb{E}_{\pi,\mathcal{M},X,Y}\left[\mathcal{R}(\mathcal{M}(X,Y))\right] = \Omega\left(\frac{d^2}{n^2\left(e^{\rho^2}-1\right)}\right). \tag{9.61}$$

Merging (9.61) with the Bayes risk obtained in Lemma J.19 gives the desired result, as the privacy cost dominates only for $\rho = O_\gamma(1)$ such that $\rho^2 = \Theta_\gamma(e^{\rho^2} - 1)$. The full argument is deferred to Appendix J.4.4.

9.6 Scaling laws

Let us now focus on the ill-conditioned setting, i.e., when the condition number of the data covariance Σ depends on the number of dimensions d. In this setting, the bounds in (9.20) are not tight enough to describe the behavior of Algorithm 9.1, as $\lambda_{\max}/\lambda_{\min}$ diverges with d.

Therefore, we analyze the full system of ODEs in (9.15) and (9.16). In particular, $D_i(t)$ has the following implicit solution

$$D_i(t) = D_i(0)e^{-2\lambda_i\Gamma(t)} + \int_0^t \left(\lambda_i\bar{\eta}^2(s)\nu_c(R(s))(R(s)+\zeta^2/2)\gamma + 2c^2\tilde{\sigma}^2(s)\gamma^2\right) e^{-2\lambda_i(\Gamma(t)-\Gamma(s))}\mathrm{d}s, \tag{9.62}$$

where we introduced the shorthand $\Gamma(t) = \int_0^t \bar{\eta}(s)\mu_c(R(s))\mathrm{d}s$. Multiplying by λ_i both sides of (9.62), averaging over $i \in [d]$, and using (9.16), we obtain the following implicit equation for the risk $R(t)$:

$$R(t) = \mathcal{F}(\Gamma(t)) + \int_0^t \left(\bar{\eta}^2(s)\nu_c(R(s))(R(s)+\zeta^2/2)\gamma\mathcal{K}(\Gamma(t)-\Gamma(s)) + 2c^2\tilde{\sigma}^2(s)\gamma^2\mathcal{J}(\Gamma(t)-\Gamma(s))\right)\mathrm{d}s, \tag{9.63}$$

where we introduced the shorthands

$$\mathcal{F}(x) := \frac{1}{d}\sum_{i=1}^d D_i(0)\lambda_i e^{-2\lambda_i x}, \qquad \mathcal{K}(x) := \frac{1}{d}\sum_{i=1}^d \lambda_i^2 e^{-2\lambda_i x}, \qquad \mathcal{J}(x) := \frac{1}{d}\sum_{i=1}^d \lambda_i e^{-2\lambda_i x}. \tag{9.64}$$

We assume a power-law distribution for the covariance spectrum.

Assumption 9.13. *Σ has a spectrum that converges weakly as $d \to \infty$ to the power law measure $p(\lambda) = (1-\phi)C_\phi^{\phi-1}\lambda^{-\phi}\mathbf{1}_{(0,C_\phi)}$, with $C_\phi = \frac{2-\phi}{1-\phi}$, for some $\phi < 1$. Furthermore, for some $\psi < 1-\phi$, we have that $D_i(0) = d\left(\omega_i^\top\theta^*\right)^2/2 = \Theta_\gamma(\lambda_i^{-\psi})$ uniformly for $i \in [d]$, with $\limsup_{d\to\infty}\sum\lambda_i^{-\psi}/d < +\infty$.*

In words, ϕ regulates the distribution of the spectrum. For $\phi = 0$, the spectrum is uniformly distributed in the interval $[0, 2]$; as $\phi \to 1$ more eigenvalues are concentrated towards 0; and as $\phi \to -\infty$, the covariance approaches the identity. The multiplicative factor C_ϕ is to ensure that the density $p(\lambda)$ is normalized and that $\mathrm{tr}(\Sigma)/d \to 1$, in agreement with Assumption 9.3. The coefficient ψ describes how the projections of θ^* on the eigenvectors ω_i of the covariance depend on the corresponding eigenvalues. For $\psi = 0$, their squares are distributed uniformly; as ψ increases, more weight is on the directions where the eigenvalues are small; and as ψ decreases, the opposite happens. The last condition is to guarantee $\|\theta^*\|_2^2 < \infty$, and it directly implies the bound $\psi < 1-\phi$.

We note that Assumption 9.13 differs from those considered in the kernel ridge regression literature, see Caponnetto and De Vito [2007], Rudi and Rosasco [2017], Defilippis et al. [2024], Paquette et al. [2024b]. Our assumption is instead inspired by Collins-Woodfin et al. [2024b], and it is better suited to the random matrix theory regime underlying Theorem 9.6, where a deterministic equivalent is derived. More precisely, existing work on kernel ridge regression sets the eigenvalues of the covariance to $\lambda_i = i^{-\alpha}$, for $\alpha > 1$, which makes $\mathrm{tr}(\Sigma)$ converge to a quantity independent of d. This is in contrast with the requirement $\mathrm{tr}(\Sigma) = d$ in Assumption 9.3, which in turn is used to prove the concentration argument in Theorem 9.6.

As in the previous section, we study the non-asymptotic behavior of $\mathcal{R}(\theta_\alpha^p)$ as $\gamma \to 0$ for the class of polynomial schedules $\tilde{\eta}(t) = \tilde{\eta}(0)(1-t)^\alpha$ with $\tilde{\eta}(0) \le 2/\gamma$. The following result (proved in Appendix J.5) considers a privacy parameter that scales polynomially with γ, i.e., $\rho = \Theta_\gamma(\gamma^b)$, and it establishes the optimal risk obtained by optimizing the hyper-parameters $(c, \tilde{\eta}(0))$ for fixed (ϕ, ψ, α, b).

Theorem 9.14. *Let Assumptions 9.3, 9.4 and 9.13 hold. Let θ_α^p be the solution of Algorithm 9.1, with $\tilde{\eta}(t) = \tilde{\eta}(0)(1-t)^\alpha$, where $\alpha = \{0, 1/2\}$ and $\alpha \geq 1$ such that $\tilde{\eta}(0) \leq 2/\gamma$. Let $\rho = \Theta_\gamma(\gamma^b)$ with $b < 1$. Denote $K = (2-\phi-\psi)(\alpha+1)$, and set $c = O_\gamma(1)$ and $c\tilde{\eta}(0) = \gamma^a$. Then,*

- *if $\phi(\alpha+1) < 2$ and $b \leq \frac{K}{2(K+1)}$, set*

$$a = -\frac{\alpha+1}{K+1}, \qquad \text{and define} \qquad h = \frac{K}{K+1}; \tag{9.65}$$

- *If $\phi(\alpha+1) < 2$ and $b > \frac{K}{2(K+1)}$, set*

$$a = -\frac{2(1-b)(\alpha+1)}{K+2}, \qquad \text{and define} \qquad h = \frac{2K(1-b)}{K+2}, \tag{9.66}$$

- *If $\phi(\alpha+1) \geq 2$ and $b \leq 1 - \frac{(2-\psi)(\alpha+1)}{2(K+1)}$, set a and define h as in (9.65);*
- *If $\phi(\alpha+1) \geq 2$ and $b > 1 - \frac{(2-\psi)(\alpha+1)}{2(K+1)}$, set*

$$a = -\frac{2(1-b)}{2-\psi}, \qquad \text{and define} \qquad h = \frac{2(2-\phi-\psi)(1-b)}{2-\psi} + \frac{\ln(a\ln\gamma)}{\ln\gamma}\,\mathbf{1}\left\{\phi(\alpha+1) = 2\right\}. \tag{9.67}$$

For all of these cases, with overwhelming probability, we have $\mathcal{R}(\theta_\alpha^p) = \Theta_\gamma(\gamma^h)$. Furthermore, for all of these cases, with overwhelming probability, we have $\mathcal{R}(\theta_\alpha^p) = \Omega_\gamma(\gamma^h)$, for all choices of c and all choices of a independent of γ.

The scaling laws of Theorem 9.14 highlight two regimes. On the one hand, if b is sufficiently small (corresponding to a sufficiently large ρ), privacy is achieved without incurring a penalty in performance; in fact, in (9.65), h does not depend on b, which implies that the final risk with optimal hyper-parameter choice is independent of the privacy requirement. On the other hand, if b exceeds a critical value (corresponding to a more stringent requirement in terms of ρ), then the cost of privacy dominates and the final risk depends on b. This is the case for both $\phi(\alpha+1) < 2$ and $\phi(\alpha+1) \geq 2$ (see respectively (9.66) and (9.67)), albeit the critical value of b and the final risk change in their dependence on (ϕ, ψ, α). We note that the critical value of b ranges within the interval $b \in (1/4, 1/2)$ for $\phi(\alpha+1) < 2$ and within the interval $b \in (0, 1/2)$ for $\phi(\alpha+1) \geq 2$. Similarly to Proposition 9.9, Theorem 9.14 shows that the optimal value of $c\tilde{\eta}(0)$ is an increasing function of $1/\gamma$, since a is always negative. Furthermore, given the admitted ranges of ϕ and ψ, we have that $a > -1$ in all cases, which guarantees the existence of $c = O_\gamma(1)$ such that $\tilde{\eta}(0) \leq 2/\gamma$.

In Figures 9.2 and 9.3, we numerically simulate the system of ODEs in 9.15 and (9.16) for different values of (ϕ, ψ, α), as γ decreases. We fix $c = 0.1$, optimize over $\tilde{\eta}(0)$, and report the minimum value of $R(1) + 2c^2\tilde{\eta}^2(1)\gamma^2/\rho^2$ for each γ. Then, we perform a linear best fit in log-log scale and report the slope as a function of α. More precisely, in Figure 9.2, we consider small values of (ϕ, ψ, α) and the privacy levels $\rho = 1$ and $\rho = \gamma^{0.5}$, which lead to the scaling laws in (9.65)-(9.66). The simulation results (circles) agree well with the values of h (stars) from Theorem 9.14, showing that our theory is predictive of the final risk with the optimal hyper-parameter choice. The plots also illustrate that h increases with K (as predicted by

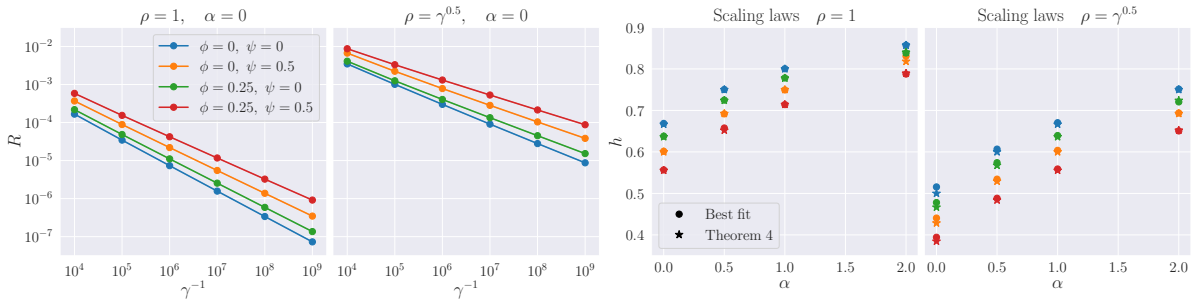


Figure 9.2: Numerical simulations for the system of ODEs in (9.15) and (9.16), for $d = 100000$, $\zeta = 0.3$ and $c = 0.1$. In the first and second panel, we report $R(1) + 2c^2\tilde{\eta}^2(1)\gamma^2/\rho^2$ for $\alpha = 0$, $\phi \in \{0, 0.25\}$, $\psi \in \{0, 0.5\}$, and $\rho \in \{1, \gamma^{0.5}\}$, after optimizing w.r.t. $\tilde{\eta}(0)$. In the third and fourth panel, we report the slope of the linear best fit in log-log scale for $\alpha \in \{0, 0.5, 1, 2\}$ (circle marker), together with the values of h given by Theorem 9.14 (star marker).

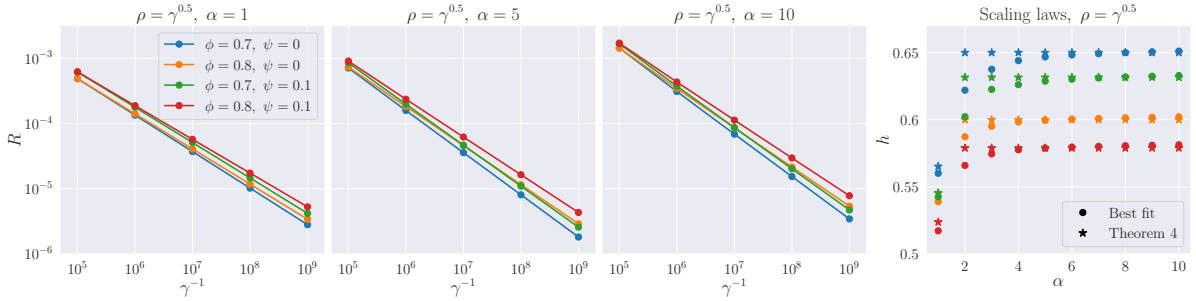


Figure 9.3: Numerical simulations for the system of ODEs in (9.15) and (9.16), for $d = 10000$, $\zeta = 0.3$ and $c = 0.1$. In the first three panels, we report $R(1) + 2c^2\tilde{\eta}^2(1)\gamma^2/\rho^2$ for $\alpha = 0$, $\phi \in \{0.7, 0.8\}$, $\psi \in \{0, 0.1\}$, $\rho = \gamma^{0.5}$, after optimizing w.r.t. $\tilde{\eta}(0)$, for different schedules with $\alpha = \{1, 5, 10\}$. In the fourth panel, we report the slope of the linear best fit in log-log scale for $\alpha \in \{1, \dots, 10\}$ (circle marker), together with the values of h given by Theorem 9.14 (star marker).

(9.65)-(9.66)), and therefore, it increases with both α and $2 - \phi - \psi$ (h increases as the color changes from red to orange, green and blue). In Figure 9.3, we consider larger values of ϕ, α and the privacy level $\rho = \gamma^{0.5}$, which lead to the scaling laws in (9.67). For values of α large enough, the simulation results (circles) agree well with the values of h (stars) from Theorem 9.14. We expect a better fit for small values of α if simulations were run for larger values of γ^{-1}, which is computationally more expensive. The plots illustrate how the value of h saturates after a certain value of α, in agreement with (9.67).

We note that, for fixed (ϕ, ψ), h is non decreasing in α, suggesting that larger values of α improve performance when γ is sufficiently small (see also the discussion after Proposition 9.10). This is formalized in the result below (also proved in Appendix J.5), which readily follows from Theorem 9.14 after taking large enough α.

Corollary 9.15. *Consider the setting of Theorem 9.14, and suppose $\phi \leq 0$. Then, for every $\epsilon > 0$, there exist α, c and $\tilde{\eta}(0)$ such that, with overwhelming probability,*

$$\mathcal{R}(\theta^p_\alpha) = O_\gamma\left(\gamma^{1-\epsilon} + \left(\frac{\gamma^2}{\rho^2}\right)^{1-\epsilon}\right). \tag{9.68}$$

Suppose instead $\phi > 0$. Then, for every $\epsilon > 0$, there exist α, c and $\tilde{\eta}(0)$ such that, with

overwhelming probability,

$$\mathcal{R}(\theta_\alpha^p) = O_\gamma\left(\gamma^{1-\epsilon} + \left(\frac{\gamma^2}{\rho^2}\right)^{\frac{2-\psi-\phi}{2-\psi}}\right). \tag{9.69}$$

We recall that the setting $\phi < 0$ is closer to the well-conditioned case discussed in Section 9.5, since the covariance approaches the identity as $\phi \to -\infty$. In fact, in this setting, the risk approaches the same rate $\Theta_\gamma(\gamma + \gamma^2/\rho^2)$ given by Theorems 9.11-9.12. This implies that the privacy requirement has a performance cost when $b \geq 1/2$, corresponding to $\rho = O(\sqrt{\gamma})$.

In contrast, $\phi > 0$ is closer to an ill-conditioned setting, since more eigenvalues concentrate towards 0 as $\phi \to 1$. In this setting, the minimax optimal rate for well-conditioned covariance is not approached. In fact, if

$$b > \frac{1}{2} - \frac{\phi}{2(2-\phi-\psi)}, \tag{9.70}$$

the second term in the RHS of (9.69) dominates the first one (when ϵ is sufficiently small). We note that the RHS of (9.70) is smaller than $1/2$. This implies that, when $\phi > 0$, the privacy requirement has a performance cost for a larger privacy parameter ρ than when $\phi < 0$. In fact, as ϕ approaches 1, even a mild privacy requirement (corresponding to b close to 0) yields a decrease in performance.

We now provide an interpretation of the higher cost of privacy incurred when $\phi > 0$. If $\phi > 0$, then Σ has many small eigenvalues, which corresponds to having several directions with low signal-to-noise ratio. In this subspace, estimating θ^* relies on rare samples with unusually large components, making the learning algorithm more dependent on atypical observations. The effect is further enhanced as more weight of θ^* is on the directions where the eigenvalues are small, which corresponds to having a larger value of ψ, as discussed after Assumption 9.13. This is reminiscent of a setting such that "learning requires memorization[4]" [Feldman, 2020], and in this scenario enforcing differential privacy has been shown to deteriorate performance especially on atypical and under-represented groups [Cummings et al., 2019, Fioretto et al., 2022].

9.7 Numerical results

Experimental details. We consider synthetic data (Figures 9.4 and 9.7), the MNIST dataset (Figures 9.5 and 9.8), and the California housing dataset (Figures 9.6 and 9.8). For experiments on synthetic data, we sample Gaussian data following Assumptions 9.3 and 9.13. For experiments on MNIST, we cast the problem as a binary task by selecting two classes (digits "1" and "7") and labeling them with $y = \pm 1$. Each image is flattened into a vector in $\mathbb{R}^d$ with $d = 784$. For experiments on the California housing dataset, we download its scikit-learn version and consider as input covariates its $d = 8$ numerical features. We split datasets into a training subset, a disjoint normalization subset, and a held-out validation/test subset. Using the normalization subset, we compute (non-privately) the empirical mean and standard deviation of each input coordinate (and of the labels) and use them to normalize both the training and validation covariates (labels), so that each coordinate has zero mean and unit variance. The plotted quantity $\mathcal{P}$ in Figures 9.5, 9.6 and 9.8 corresponds to the average value of the final validation square loss. For the housing dataset, we notice that our

[4]Memorization captures the influence an individual sample has on the final model.

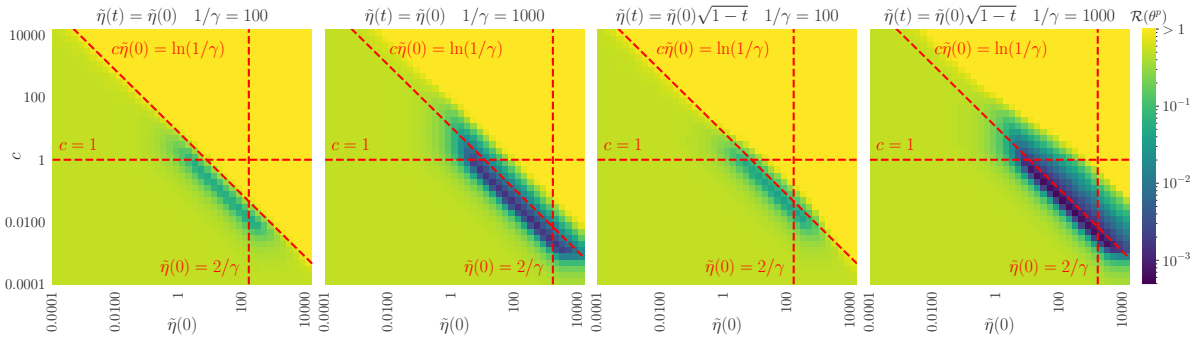


Figure 9.4: Numerical simulations for $\mathcal{R}(\theta^p)$ obtained via Algorithm 9.1 for $d = 100$ and $\Sigma = I$, as a function of c and $\tilde{\eta}(0)$. We consider the schedules corresponding to output perturbation ($\tilde{\eta}(t) = \tilde{\eta}(0)$, first and second panel) and constant private noise ($\tilde{\eta}(t) = \tilde{\eta}(0)\sqrt{1-t}$, third and fourth panel), with fixed $\rho = 0.1$ and $\zeta = 0.3$. We set $\gamma = 0.01$ in the first and third panel, and $\gamma = 0.001$ in the second and fourth panel. The values of $\mathcal{R}(\theta^p)$ are capped at 1, and θ^* is chosen such that $\mathcal{R}(\theta_0) = 0.5$. We indicate with red dashed lines the curves $c = 1$, $\tilde{\eta}(0) = 2/\gamma$, and $c\tilde{\eta}(0) = \ln(1/\gamma)$, and we display the average over 10 independent trials.

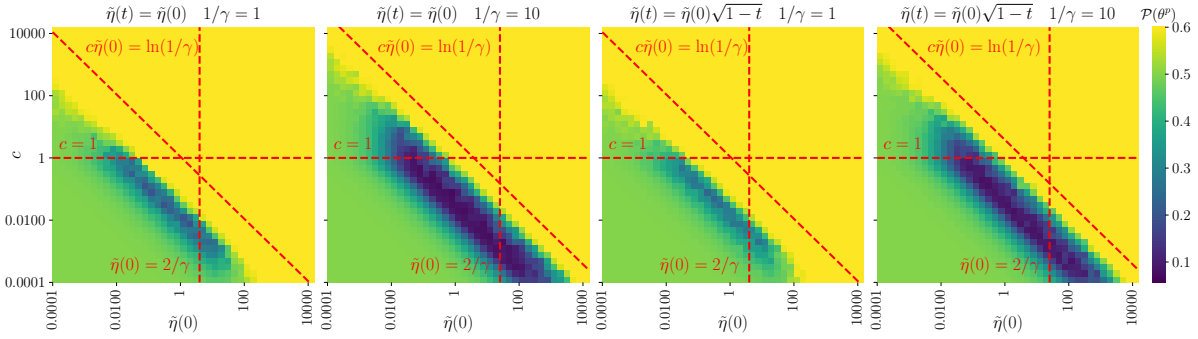


Figure 9.5: Numerical simulations for $\mathcal{P}(\theta^p)$ obtained via Algorithm 9.1 on the MNIST dataset, as a function of c and $\tilde{\eta}(0)$. We consider the schedules corresponding to output perturbation ($\tilde{\eta}(t) = \tilde{\eta}(0)$, first and second panel) and constant private noise ($\tilde{\eta}(t) = \tilde{\eta}(0)\sqrt{1-t}$, third and fourth panel), with fixed $\rho = 0.1$. We set $\gamma = 1$ in the first and third panel, and $\gamma = 0.1$ in the second and fourth panel. The values of $\mathcal{P}(\theta^p)$ are capped at 0.6. We indicate with red dashed lines the curves $c = 1$, $\tilde{\eta}(0) = 2/\gamma$, and $c\tilde{\eta}(0) = \ln(1/\gamma)$, and we display the average over 10 independent trials.

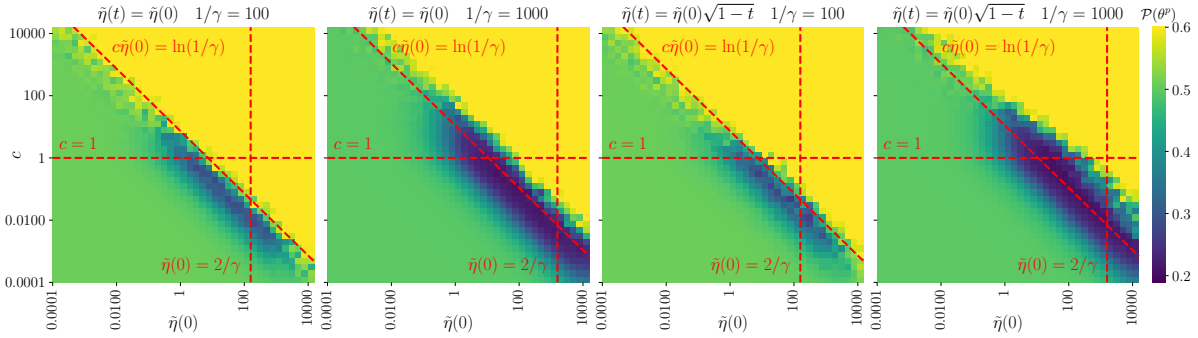


Figure 9.6: Numerical simulations for $\mathcal{P}(\theta^p)$ obtained via Algorithm 9.1 on the California housing dataset, as a function of c and $\tilde{\eta}(0)$. We consider the schedules corresponding to output perturbation ($\tilde{\eta}(t) = \tilde{\eta}(0)$, first and second panel) and constant private noise ($\tilde{\eta}(t) = \tilde{\eta}(0)\sqrt{1-t}$, third and fourth panel), with fixed $\rho = 0.1$. We set $\gamma = 0.01$ in the first and third panel, and $\gamma = 0.001$ in the second and fourth panel. The values of $\mathcal{P}(\theta^p)$ are capped at 0.6. We indicate with red dashed lines the curves $c = 1$, $\tilde{\eta}(0) = 2/\gamma$, and $c\tilde{\eta}(0) = \ln(1/\gamma)$, and we display the average over 5 independent trials.

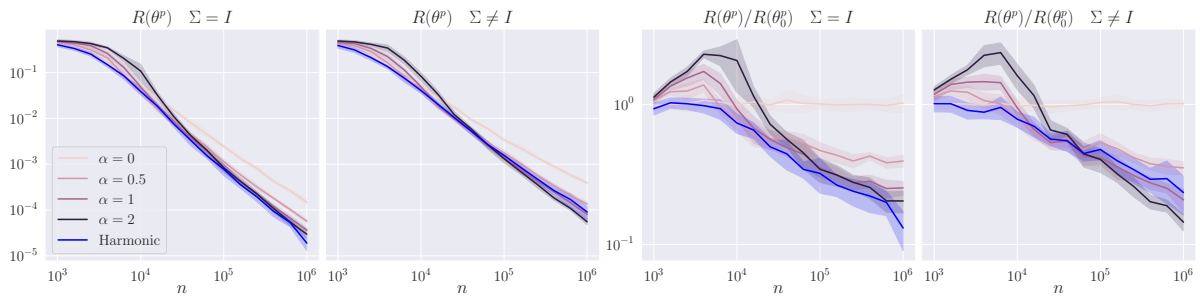


Figure 9.7: Numerical simulations for $\mathcal{R}(\theta^p)$ obtained via Algorithm 9.1 for $d = 100$, $\zeta = 0.3$, $(\theta_i^*)^2 = 1/d$, and $\rho = 0.1$, as a function of n. We fix $c = 0.1$ and consider the polynomial schedules in (9.33) for $\alpha \in \{0, 0.5, 1, 2\}$ (optimizing w.r.t. $\tilde{\eta}(0)$) and the harmonic schedule in (9.50) (optimizing w.r.t. β and τ). We report the average over 10 independent trials, as well as the confidence interval corresponding to 1 standard deviation. We consider both isotropic data ($\Sigma = I$, first and third panel) and data with diagonal covariance whose eigenvalues are uniformly distributed in the interval $[0, 2]$ (second and fourth panel).

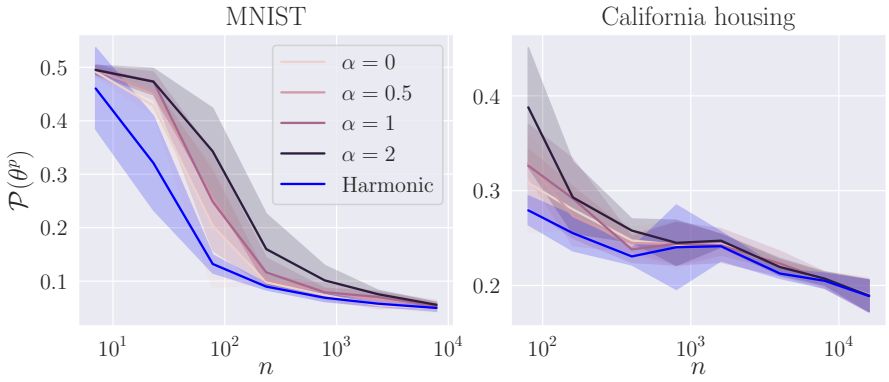


Figure 9.8: Numerical simulations for $\mathcal{P}(\theta^p)$ obtained via Algorithm 9.1 with $\rho = 1$ for the MNIST and the California housing dataset as a function of n. The method is in common with the one in Figure 9.7.

algorithm becomes numerically unstable for larger values of n, with the value of the final loss being disproportionally large for a few values of the random seeds. We manually exclude such seeds from the plots.

Hyper-parameter space. In Figures 9.4, 9.5 and 9.6, we run Algorithm 9.1 with a fixed privacy budget $\rho = 0.1$ on synthetic, MNIST and California housing data. We focus on the schedules $\tilde{\eta}(t) = \tilde{\eta}(0)$ and $\tilde{\eta}(t) = \tilde{\eta}(0)\sqrt{1-t}$, for varying values of the renormalized clipping constant c and learning rate $\tilde{\eta}(0)$. We report on the heatmaps the resulting values of the final loss as a function of $\tilde{\eta}(0)$ (on the x axis) and c (on the y axis). All the results are in agreement with our discussion on the hyper-parameters following Proposition 9.9: performance deteriorates if either c exceeds 1 (upper part of the heatmaps) or $\tilde{\eta}(0)$ exceeds $2/\gamma$ (right part of the heatmaps), and the lowest values of the risk are roughly parallel to the line $c\tilde{\eta}(0) = \ln(1/\gamma)$. We remark that these scalings agree with the empirical practice in deep learning of using a sufficiently small clipping constant, with a learning rate renormalized by its value [De et al., 2022, McKenna et al., 2025].

Effect of the learning rate schedule. In Figures 9.7 and 9.8, we compare different schedules for the learning rate $\tilde{\eta}(t)$, after setting $c = 0.1$ and optimizing $\tilde{\eta}(0)$ numerically. We consider the polynomial schedules in (9.33) for $\alpha = \{0, 0.5, 1, 2\}$ and the harmonic schedule in (9.50). The first and second panel in Figure 9.7 (which display the results for synthetic data) show that all schedules rapidly give better results as n increases, but larger values of α are optimal only for large enough n. This effect is clearly shown in the third and fourth panel, which reports the same results normalized by the loss of output perturbation ($\alpha = 0$).

Hence, if n is sufficiently small compared to d, output perturbation can outperform other schedules, in agreement with an observation made in Dwork et al. [2025] regarding the comparison between output and objective perturbation. Furthermore, the benefits of using larger values of α for sufficiently large values of n seem more apparent in the case of non-isotropic covariance, displayed in the second and fourth panel. The results for the harmonic schedule (reported in blue) are obtained after optimizing numerically β and τ as in (9.50). The third panel of Figure 9.7 clearly displays the benefits of this choice for both small and large values of n. In the non-isotropic case (fourth panel), this schedule performs optimally for sufficiently small values of n, and it is on par with other polynomial schedules when $n = 10^4 d$ (right part of the plot). In Figure 9.8, we run the same experiment for the MNIST and California housing datasets: we note that the harmonic schedule remains optimal, as suggested by our theory.

9.8 Discussion and future directions

We show optimal rates for DP linear regression in the proportional regime where the number of samples grows linearly with the input dimension. To achieve this result, a crucial technical innovation is to establish a deterministic equivalent of the test risk of DP-GD via a family of deterministic ODEs, in the challenging setting where the clipping constant is of the order of the per-sample gradient. Analyzing the family of ODEs then gives bounds that are sharp enough to *(i)* demonstrate the benefit of small clipping constants, and to *(ii)* identify the optimal learning rate schedule.

The optimality of the rate $\gamma + \gamma^2/\rho^2$ holds for a well-conditioned covariance, as shown in the minimax lower bound of Theorem 9.12. In the ill-conditioned setting, we consider a covariance spectrum distributed according to a power law (Assumption 9.13), and we provide a tight analysis of the DP-GD risk for polynomially-decaying learning rate schedules. As in the well-conditioned case, our results highlight the benefits of small clipping constants and decaying schedules. However, in contrast to the well-conditioned case, establishing minimax rates remains an interesting direction for future work. We also note that the optimality of our rates holds when the variance of the label noise ζ^2 is a strictly positive constant independent of γ. We cautiously suspect that, for $\zeta = o_\gamma(1)$, the optimal rate is of order $\zeta^2(\gamma + \gamma^2/\rho^2)$ (see e.g. Theorem 4.1 in Cai et al. [2021], where this dependence is tracked). In this regime, our bounds on $\mu_c(R)$ and $\nu_c(R)$ in Lemma J.4 suggest that a sufficiently small value of the clipping constant is $c = O_\gamma(\sqrt{2R + \zeta^2})$ rather than $c = O_\gamma(1)$. If $\zeta = o_\gamma(\gamma)$, setting a fixed $c = O_\gamma(\zeta)$ might not allow to achieve the optimal rates (as in the right side of the panels in Figure 9.4), suggesting the need for a time-dependent scheduling of c. Interesting future directions also involve *(i)* proving that the optimal non-asymptotic rates read $d/n + d^2/(n\rho)^2$ beyond the proportional regime, and *(ii)* better understanding the precise proportional asymptotics beyond the limit $\gamma \to 0$, for example showing the benefits of output perturbation w.r.t. schedules with

decaying learning rate when γ is sufficiently large, as displayed in Figure 9.7.

Finally, we remark that the technical approach of analyzing DP-GD through a deterministic equivalent can be used to study differentially private optimization beyond linear regression. A concrete setting for future work is provided e.g. by logistic regression, where the GD dynamics (in the absence of clipping and private noise) has been considered in Collins-Woodfin and Seroussi [2025].

CHAPTER 10

Discussion and Future Directions

In this thesis, we study the asymptotic behavior of machine learning models, such as over-parameterized neural networks and linear models in the regime of high-dimensional data, trained with either empirical risk minimization or differentially private gradient descent. Then, the main focus is on the properties of the trained models in the context of data memorization, input-output robustness, and privacy-utility trade-offs. In particular, we study how these properties depend on the scalings of the problem, namely the number of training samples n, the input dimension d, and the number of parameters p.

Thesis Summary and Contributions

In Chapters 2, 3, and 4, we focus on the problem of memorization in over-parameterized models. In Chapter 2, we show that a fully-connected deep neural network can fit n labels with gradient descent as long as its number of parameters is $p = \tilde{\Omega}(n)$, which achieves the information-theoretic lower bound for this problem. The proof technique relies on providing a tight lower bound on the smallest eigenvalue of the empirical neural tangent kernel of the network at initialization. In Chapter 3, we focus on the problem of spurious feature memorization, where each training sample has two components $[x, y]$, where y is uninformative for the task, but is nevertheless memorized by the model, and can be used at test time to infer information about the corresponding core feature x. The techniques used in this chapter rely on quantifying the stability of over-parameterized generalized linear models to leave-one-out perturbations of the training set, which is a useful step also for the proof of the results in Chapter 8. Then, in Chapter 4, we consider the problem of reconstructing the full training dataset from the model parameters, focusing on an over-parameterized RF model. There, we give evidence for the threshold $p = \Omega(dn)$ for possible data reconstruction, which (in a high-dimensional setting) is much larger than the one for label fitting, $p = \Omega(n)$, from Chapter 2. The proof techniques are similar to standard analyses of the RF kernel concentration, but require uniform guarantees in data space to ensure that the only solution of our reconstruction algorithm in the dataset space is indeed the original training set. Such uniform guarantees are obtained via packing arguments on the d-dimensional sphere, which translate into an extra factor of d in the typical lower bound $p = \tilde{\Omega}(n)$ required to achieve RF kernel concentration.

In Chapters 5, 6, and 7, we discuss problems connected to robustness in machine learning. In Chapter 5, we focus on interpolating over-parameterized models and study their robustness

to adversarial perturbations, in a spirit similar to the previous work of Bubeck et al. [2021]. There, we show that some learning models are non-robust no matter their degree of over-parameterization, while others can be robust if their number of parameters is at least $p = \Omega(dn)$. This second family also includes some appropriately initialized two-layer neural networks, thus resolving in the affirmative a previous conjecture from Bubeck et al. [2020]. Notably, this scaling matches the one discussed in Chapter 4 for data reconstruction. We expect this connection to be not just a matter of coincidence, as both problems are strictly tied to the behavior of input space gradients of the model: their norm characterizes the sensitivity to adversarial perturbations, and they are used to optimize the reconstruction loss that recovers the original training images in Chapter 4. In Chapter 6, we instead focus on the problem of spurious correlations, where the training data are divided into two components $[x, y]$, where x is the core feature, which is informative for the task, and y is a spurious feature, which is (differently from the setting analyzed in Chapter 3) correlated with x, and therefore with the label. There, we initially focus on high-dimensional linear regression, where $n = \Theta(d)$, and we rely on the asymptotic analysis of Han and Xu [2023] to characterize the test error and the amount of spurious correlations learned by the model as a function of the weight decay parameter. Then, through an equivalence argument between linear regression and RF regression, we connect these findings to the impact of over-parameterization for this same problem. Finally, in Chapter 7, we focus on the study of the word sensitivity of attention-based models, informally meant as how much the output of the model changes when a single word in the input sequence is changed. In this context, we frame word sensitivity as a desirable property for a language model, as contextual meaning in a sentence is often tied to the presence of a specific word. Then, we show that, while the word sensitivity of an RF model vanishes with the length of the context, it remains lower bounded by a constant for randomly initialized attention layers, as long as the embedding dimension is of at least the same order as the context length.

Finally, in Chapters 8 and 9, we study differentially private optimization, both in an over-parameterized RF model, and for DP linear regression. In Chapter 8, we challenge the common wisdom that over-parameterization inherently hinders performance in private learning, showing that a DP-GD algorithm with proper hyper-parameters can achieve the same asymptotic test error as non-private GD in an arbitrarily over-parameterized RF model. More at a high level, we argue that the reason why other upper bounds on the test risk of DP algorithms in the over-parameterized regime become vacuous is that they do not capture the implicit bias of DP gradient descent, but rather rely on uniform stability arguments, which have been argued to be unsuitable for capturing the generalization properties of gradient descent in the over-parameterized regime even in the non-private setting Zhang et al. [2017b]. In Chapter 9, we instead focus on the problem of DP linear regression in the high-dimensional proportional regime $n = \Theta(d)$. There, we show that, in the well-conditioned case, a DP-GD algorithm can achieve the minimax rates for this task, highlighting the importance of sufficiently small gradient clipping constants, which allow one to reduce the amount of private noise added at each iteration. This is in contrast with existing nearly optimal DP-GD results for linear regression, which typically set the clipping constant large enough that clipping almost never occurs Varshney et al. [2022], Liu et al. [2023b], Brown et al. [2024b], allowing for a simpler analysis. Our method is inspired by Marshall et al. [2025], and relies on comparing the test risk of DP-GD to the solution of a deterministic system of ordinary differential equations, which in turn allows us to tightly analyze the dynamics of the algorithm even when clipping is done aggressively.

Future Directions

The problems discussed in this thesis open a number of related research directions connected to topics in data memorization and private optimization.

Data memorization. First, we have yet to fully understand the phenomenology described in Chapter 4, i.e. the role of the model size for data reconstruction. Even in the simplified setting of a RF model, it is still unclear which are the precise settings under which sufficiently small models provably do not memorize, whereas sufficiently large models provably do. This would ideally result in identifying the threshold (possibly $p = \Theta(dn)$) that separates information-theoretic impossibility from computational feasibility of data reconstruction. For the information-theoretic impossibility, we remark the importance of proper assumptions on the input training data distribution, such as a notion of anti-concentration or effective dimensionality. In fact, if data are restricted to a d'-dimensional subspace of $\mathbb{R}^d$, we expect $p = \Omega(d'n)$ parameters to contain enough information to enable reconstruction, at least in a setting comparable to that of Chapter 4. Importantly, all these results depend on the precise definition of memorization. In Chapter 4, this is associated with the possibility of reconstructing any training sample *without* any knowledge of the remaining part of the dataset. In fact, we expect an informed and powerful adversary with knowledge of the rest of the dataset to be able to reconstruct the missing one even in the regime $p = \Omega(d)$ [Balle et al., 2022]. Next, we consider an exciting avenue for future research the problem of understanding the interplay and trade-offs between reconstruction and the privacy parameter ρ (see Definition 9.1) in a DP-trained model. We expect this problem to reveal a phase transition at $\rho = \Theta(\sqrt{d})$ [Balle et al., 2022], highlighting the relevance of studying DP optimization algorithms in the regime where the privacy budget is much larger (as it is, in fact, "high-dimensional") than the usual constant-order values more commonly used in practice. This privacy scaling could allow to train learning models that are robust against data reconstruction attacks, with a smaller penalty in performance.

Private optimization. In Chapter 9 we achieved the minimax rates for the task of DP linear regression. However, this result is limited to the proportional regime $n = \Theta(d)$ and to the case where the label noise is of constant order. We expect it to be possible to improve on these limitations, perhaps by focusing on algorithms that could offer an easier theoretical analysis, such as output perturbation as in Dwork et al. [2025] analyzed with the convex Gaussian min-max theorem [Thrampoulidis et al., 2015]. Beyond linear regression, we expect analyses based on concentration to deterministic equivalents (see, e.g., Collins-Woodfin and Seroussi [2025]) to allow to achieve progress in our understanding of *DP logistic regression* with standard gradient methods (such as Algorithm 9.1), without the need of any dimensionality reduction techniques [Bassily et al., 2022, Wang et al., 2025]. Both these directions have the implicit goal of characterizing the role of *hyper-parameter optimization* in practical DP algorithms, such as the benefit of using sufficiently small clipping constants.

The cost of $\rho = \Omega(\sqrt{d})$. A separate direction is connected to studying the utility costs of privacy budgets larger than $\rho = \Theta(\sqrt{d})$, which we argued in a previous paragraph is relevant in the context of data reconstruction. Let us consider the problem of linear regression, discussed in Chapter 9. There, we showed that, in the well-conditioned and proportional regime, the minimax rates for this task, informally denoted by $\mathcal{R}$, are

$$\mathcal{R} = \Theta\left(\frac{d}{n} + \frac{d^2}{n^2\rho^2}\right).$$

If we suppose these rates to hold on a wider set of scalings, and we set $\rho = \Omega(\sqrt{d})$, we get $\mathcal{R} = \Theta(d/n + d/n^2) = \Theta(d/n)$, which is the same rate as in the non-private case. This is a very preliminary calculation, but it suggests that in the regime of privacy budgets larger than our guessed data reconstruction threshold, $\rho = \Omega(\sqrt{d})$, the cost of privacy is always negligible relative to the non-private statistical error. Intuitively, this could follow from the fact that fully memorizing the input data should not provide any statistical advantage in learning patterns shared across data. More generally, we believe that, for a wider class of statistical tasks (such as linear regression, logistic regression, single-index models, etc.), this regime may be proven to always allow one to achieve *privacy for free*. We consider this an interesting future direction for making progress in our general understanding of the fundamental cost of privacy in different settings, and its relation to the Bayes risk.

The cost of the worst-case. Another research direction is associated with understanding the cost of the worst-case nature of DP, which by definition provides stability guarantees uniformly over any pair of adjacent datasets. This requirement is sometimes referred to as too strict, and potentially responsible for the utility losses often associated with private optimization. For this reason, several relaxations have been proposed to capture more distributional or average-case notions of privacy, such as *random DP* [Hall et al., 2013] or *Bayesian DP* [Triastcyn and Faltings, 2020]. These works are perhaps an evidence of a pessimism around calibrating algorithms to worst-case, pathological, or statistically irrelevant inputs. Specifically to the material covered in this Thesis, we highlight that in the proof of the lower bound for DP linear regression in Chapter 9, the definition of zCDP is only used across two datasets with i.i.d. Gaussian data (see (9.58)), and this is sufficient to provide a matching lower bound (up to numerical constants) to the upper bound achieved by DP-GD, which in contrast guarantees privacy in the worst case. Then, we raise the following research questions: in which settings is *the cost of worst-case* negligible? In other words, in which settings can a DP algorithm match the same performance as algorithms satisfying a relaxed notion of privacy, such as random or Bayesian DP? In which other settings is the cost of worst-case, instead, significant? Importantly, while privacy is guaranteed in the worst case, we consider the utility of a DP algorithm *in distribution*, and determined by the particular statistical model, as done in Chapters 9 and 8, and as commonly done in prior work on private learning [Bassily et al., 2014, Jain and Thakurta, 2014, Bassily et al., 2019, Song et al., 2021b].

Copyright in machine learning. Due to the decrease in accuracy and computational overhead of DP optimization algorithms, modern foundation models are generally pre-trained without DP. This in turn makes it possible to design prompts that allow training *data extraction*, meant as data reconstruction solely through querying the model, and this evidence has led to lawsuits against AI companies for copyright infringement [Pope, 2024, Barnes, 2025]. These lawsuits are generally associated with the trained model *storing* some of the training (proprietary) data, which highlights the importance of bridging our understanding of data memorization with concepts in copyright law, generally intended to protect the *original expression* of creative works, but not the *ideas* or *facts* they contain [U.S. Copyright Office, 2025]. Recent work has attempted to build such a connection [Cooper and Grimmelmann, 2024, Vyas et al., 2023, Elkin-Koren et al., 2024, Cohen, 2025], albeit with different – and sometimes conflicting – perspectives. The high-dimensional nature of the creative works typically involved in these lawsuits (snippets of hundreds of words from newspaper articles reconstructed "near-verbatim" [Pope, 2024]) suggests that the relevant threat model involves an *attacker* who reconstructs a number of bits growing with the ambient dimension $\Theta(d)$ of

a protected work. This contrasts with the binary *membership inference* attacks commonly considered in privacy research, which suggests that the high-dimensional regime typically considered in this Thesis offers the ideal framework for studying this problem. Tackling this problem analytically may appear in tension with the inherently interpretive nature of copyright law, but we instead redeem it logically coherent to how we now (quantitatively) encode user privacy in a single real parameter ρ. Some concrete questions, from the *defendant*'s perspective, are: *(i)* in which settings is it sufficient to train the model with a privacy parameter satisfying $\rho = o(\sqrt{d})$, such that *none* of the training data can be reconstructed, while incurring only negligible statistical costs from DP? *(ii)* In practice, would training a foundation model with $\rho = \Theta(\sqrt{d})$ significantly improve robustness against current extraction methods, while maintaining acceptable performance? [1] Some other questions, from the *plaintiff*'s perspective, are the following: *(i)* in which settings is it possible to provide indisputable proof that a model has "copied" the training data? One open research direction is to assess the effectiveness of *watermarks* [Wei et al., 2024, Rastogi et al., 2025] embedded in the original data: if these are reconstructed, the result cannot be attributed to the model learning generalizable patterns, since the watermark is (statistically) unique. Then, *(ii)* in which settings would this method be robust, and how can it be made quantitative beyond membership inference?

Beyond the over-parameterized regime

A substantial amount of work in this thesis concerns the over-parameterized regime, meant as the regime where learning models have a number of trainable parameters larger than the number of training samples, and implicitly meant as a regime where the models overfit the training data until they achieve zero training error (with the exception of Chapter 8). Notably, this regime is *not* a good representation of how foundation models or large language models are trained at the time this thesis is being written. This is because *(i)* the benefit of overfitting is unclear when training auto-regressive generative language models, and *(ii)* even if there were a benefit, the computational costs of training such models have shifted the paradigm toward developers making only one or a few passes over the vast amount of data available on the web.

Given this state of affairs, we highlight the research direction of understanding the asymptotic behavior of machine learning models in the regime where a single pass over the training data is performed, hence far from the interpolating regime. We expect this regime to be qualitatively very different in the context of memorization and data reconstruction, as every training sample is seen only once, and therefore, at least intuitively, is less likely to be memorized by the model. On the other hand, there is empirical evidence that training data can be reconstructed from the parameters of language models [Carlini et al., 2021], which suggests that this phenomenon is still relevant and theoretically observable as well.

One-pass non-convex private optimization. In this direction, an exciting and challenging problem is that of differentially private pre-training of foundation models. In fact, privacy accounting methods are typically not optimal in the regime where the batch size is much smaller than the size of the full training set [Nasr et al., 2025b], making the noise of private optimization in this regime perhaps too large and overly conservative. For convex optimization, this problem can be avoided with accounting methods based on privacy amplification by iteration Feldman et al. [2018], as done in Chapter 9, where the minimax rates are achieved even with batch size 1. In non-convex deep learning optimization, a promising approach to

[1] Importantly, this question raises the problem of identifying the effective value of d for real-world data.

avoid the problem of sub-optimal privacy accounting is to rely on the recently proposed DP *follow the regularized leader* algorithm Kairouz et al. [2021], which introduces correlated noise at every iteration. This method guarantees the privacy of every single iterate, unlike privacy amplification by iteration, but the self-canceling nature of the noise allows for better utility at the level of the final iterate. This is a very preliminary research idea, and its computational and technical feasibility must be further investigated.

Post training and reinforcement learning. Beyond pre-training, post-training methods are increasingly emerging as powerful approaches to improve the performance of AI systems on complex tasks. Among these, we identify test-time compute [OpenAI, 2024, Snell et al., 2025] and reinforcement learning [Christiano et al., 2017, Ouyang et al., 2022] as two interesting avenues for future research. In particular, the success of these methods depends on scaling variables that are not considered in this thesis, such as the number of parallel reasoning branches in best-of-N test-time scaling [Brown et al., 2024a, Snell et al., 2025] (in Brown et al. [2024a], sampling up to $N = 10^4$ times per problem is shown to significantly boost performance on math and coding tasks), the amount of time spent on chain-of-thought reasoning [Wei et al., 2022, OpenAI, 2024], and the number of roll-outs (or group size) in reinforcement learning policy optimization [Schulman et al., 2017, Shao et al., 2024]. A theoretical analysis of these methods would benefit from a simplified parametric and data model able to capture the key statistical ingredients of these modern learning paradigms, in the same way that the RF model provides a useful prototype for studying the over-parameterized regime in supervised learning. We discuss in the next paragraph a preliminary idea in the direction of identifying and analyzing such a model.

A linear model for policy optimization. Consider the supervised learning problem where we have n training samples $\{(x_i, y_i)\}_{i=1}^n$, such that $x_i \in \mathbb{R}^{L\times d}$ and $y_i \in \mathbb{R}^d$. Intuitively, it is useful to think each x_i to be a *question*, made of L tokens in dimension d. Each y_i will be the corresponding *answer* to the question, which we can think as a single token. Consider the following data model

$$z_{i(t+1)} = A_i z_{it} + v_i,$$

and

$$x_{it} = z_{it} + \xi_{it}, \qquad y_i = z_{iT^*} + \xi_{iT^*}, \quad \text{with} \quad T^* > L.$$

This model lets every training question-answer i be generated recursively over t, according to the linear map described by A_i and v_i. The recursion is defined over some noiseless variables z_{it}, and each token and answer contains some noise ξ_{it}. T^* describes how many steps in time the label is separated from the questions. Intuitively, if $T^* = L + 1$ we are looking at the case where the answer follows "directly after" the question, with the problem becoming more challenging as T^* grows larger. Then, consider the following linear, auto-regressive, parametric model

$$f(\theta, x) = \left[x_{[-l:]}\right]^\top \theta,$$

where $\theta \in \mathbb{R}^l$, $x_{[-l:]} \in \mathbb{R}^{l\times d}$ denotes the matrix made by the last l rows (tokens) of x, with $l \leq L$. This choice implies that the predictive model has a context window of l and its prediction only depends on the last l tokens. Then, we can define the next token generation as

$$x_{\text{next}} \sim p\left(x_{\text{next}}|x;\theta\right) = \mathcal{N}\left(f(\theta, x), \sigma^2 I_d\right).$$

At this point, we can train this parametric model with *Reinforce* [Williams, 1992]. This requires the definition of a reward function. To do so, we fix a number of tokens T that defines the

length of the final context (question plus answer text). In other words, the answer to the question will be in the form of $T - L$ tokens, and the reward will be computed on the very last token. We fix the reward to be the negative squared distance between the label y and the token generated after $T - L$ iterations, *i.e.*

$$r(y, x_{1:T}) = - \left\| y - x_T \right\|_2^2 ,$$

where $x_{1:T} \in \mathbb{R}^{T \times d}$ denotes the full sequence composed of the original question $x_{1:L}$, and the tokens generated by the model $x_{L+1:T}$. Then, we train θ to maximize the (empirical) reward. This can be written as

$$\mathcal{R}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{p\left(x_{i,[L+1:T]} | x_{i,[1:L]}; \theta\right)} \left[r(y_i, x_{iT}) \right] . \tag{10.1}$$

This reward is in practice optimized via gradient ascent, estimating the expectation via multiple roll-outs of the generative model over the training samples. In this setup, we can effectively write the expectation above in closed form (which corresponds to the case of infinite roll-outs), and optimize this directly. Then, we leave as future work the problem of understanding the behavior of the final reward of the trained model, as a function of the scalings of the problem, and of the parameters L, T, and T^*, which could highlight the benefits of intermediate reasoning (larger T) before giving the final answer.

A broader perspective on ML research

The discussion in this final chapter has been mainly technical and specific to the work performed in the past years, and to the research projects that could be meaningful to pursue in the next few years. What follows is a more high-level discussion of the broader role of a theoretician working in the field of AI in this decade, both within the research community and, more generally, in society. The following discussion will be written in the first person, as it will necessarily contain personal opinions and reflections.

AI as an empirical science. AI and ML are evolving faster than ever, and in the past few years the massive economic consequences of AI dominance have further shifted the research culture toward "building things", rather than slowing down and taking the time to understand and inspect existing technology. Possibly, this trend has also affected the academic culture in this field, where hitting the next deadline and having a lot of papers on what is trendy seem to matter more than producing valuable science. This environment is unfortunately not very fertile for the goal of building deep and long-standing theory, able to capture and conceptualize concrete phenomena and to meaningfully translate them into a mathematical language. Furthermore, much of the published theory of learning in the past decade does not seem to properly capture some behaviors of neural networks, or is arguably restricted to highly specific settings (2-layer neural networks, Gaussian distributions, ...). Overall, a common criticism is that learning theory has not had any direct implication for AI development, but instead has been limited to identifying patterns visible from empirical evidence, and engineering specific assumptions and settings that would lead to such patterns. In this sense, deep learning theory was decently successful in its ability to *describe* some empirical phenomena in deep learning, but was quite unsuccessful in *predicting* other phenomena a priori, or in helping with the *design* of better algorithms.

The role of theory for AI practice. Given the premise of the previous paragraph, it is natural to question the role and the value of theoretical research in the field of AI. At a high level, a common argument in favour of any theoretical research is that even if the original results turn out to have limited direct practical relevance, the concepts and tools developed to achieve them may not. It is in fact hard to predict a priori which ideas will eventually become useful in the long term, and while this may generally be unlikely, theoretical research does not require large amounts of resources (such as compute, lab equipment, or large teams), making it relatively cost-effective.

In the context of machine learning, we are at a stage where the lack of transparency of large models limits much of the value that we can extract from them, especially in high-stakes applications. Thus, I would argue that theoretical effort aimed at clarifying failure modes and unpredictable behaviors of large systems is something the whole community benefits from. Such theory does not need to exactly characterize the precise behavior of large, complex architectures, but should help experts (including non-theoreticians) gain insight into general phenomena.

Furthermore, modern training methods tend to be a broad collage of different engineering steps and methods, with multiple moving parts. An example close to this thesis is differentially private training, which requires gradient clipping, noise injection, and early stopping (together with the choice of potentially different optimizers, subsampling ratios, etc.). This high degree of complexity generates many confounding variables, making it hard to understand the role of each component from experiments alone. Ideally, theory should help with the successful implementation of such algorithms (e.g. proving that aggressive clipping is beneficial, see Chapter 9), turning training from trial-and-error engineering into a more straightforward path, without the risk of running into a combinatorial explosion of ablations, which is not affordable in a field where the scale of experiments does not seem to stop growing.

Of course, theory should not be seen only as a way to boost experimental work. Its role is also to conceptualize and formalize problems that we know are important, but do not yet know how to express precisely. I believe differential privacy is a success story in this direction, as it connected statistical and cryptographic perspectives within a mathematical framework that has remained solid and influential twenty years later. This has allowed us to turn a societal need into a mathematical quantity that can be studied, optimized, and formally debated.

Our role in society. The healthy development of a solid science of AI matters to society as a whole, including to people far from the technical community. Indirectly, moving the field in a meaningful direction, educating students to aim for high standards of scientific practice, and promoting transparency can and will have broad consequences. Still, this does not exempt academics from asking what direct contribution they can make to the broader ecosystem. Academic work offers an exceptional degree of intellectual freedom, and this freedom should also motivate us to pursue directions with a concrete societal impact. I believe that the coming decades will be shaped by fields at the intersection of AI and policy, education, healthcare, and science. Technical experts will play a role in shaping this future, and I acknowledge my responsibility for what is to come.

Bibliography

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *ACM SIGSAC Conference on Computer and Communications Security*, 2016.

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016): 493–500, 2024.

Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. Differentially private assouad, fano, and le cam. In *Algorithmic Learning Theory*, pages 48–78. PMLR, 2021.

Radoslaw Adamczak. A note on the Hanson-Wright inequality for random vectors with dependencies. *Electronic Communications in Probability*, 20:1–13, 2015.

Radoslaw Adamczak, Alexander E Litvak, Alain Pajor, and Nicole Tomczak-Jaegermann. Restricted isometry property of matrices with independent columns and neighborly polytopes by random sampling. *Constructive Approximation*, 34(1):61–88, 2011.

Ben Adlam and Jeffrey Pennington. The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization. In *International Conference on Machine Learning*, 2020.

Ben Adlam, Jake Levinson, and Jeffrey Pennington. A random matrix perspective on mixtures of nonlinearities for deep learning. *arXiv preprint arXiv:1912.00827*, 2019.

Robert J. Adler and Jonathan E. Taylor. *Random Fields and Geometry*, volume 1 of *Springer Monographs in Mathematics*. Springer New York, NY, 2007. ISBN 978-0-387-48112-8.

Faruk Ahmed, Yoshua Bengio, Harm van Seijen, and Aaron Courville. Systematic generalisation with group invariant predictions. In *International Conference on Learning Representations*, 2021.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, 2019.

Kareem Amin, Alex Kulesza, Andres Munoz, and Sergei Vassilvtiskii. Bounding user contributions: A bias-variance trade-off in differential privacy. In *International Conference on Machine Learning*, 2019.

Prashanti Anderson, Ainesh Bakshi, Mahbod Majid, and Stefan Tiegel. Sample-optimal private regression in polynomial time. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, pages 2341–2349, 2025.

Galen Andrew, Om Thakkar, Hugh Brendan McMahan, and Swaroop Ramaswamy. Differentially private learning with adaptive clipping. In *Advances in Neural Information Processing Systems*, 2021.

Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, et al. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2020.

Raman Arora, Raef Bassily, Cristóbal A Guzmán, Michael Menart, and Enayat Ullah. Differentially private generalized linear models revisited. In *Advances in Neural Information Processing Systems*, 2022.

Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, 2019a.

Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, Ruslan Salakhutdinov, and Rousong Wang. On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems*, 2019b.

Devansh Arpit, Stanisław Jastrzundefinedbski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In *International Conference on Machine Learning*, 2017.

Hilal Asi, John Duchi, Alireza Fallah, Omid Javidbakht, and Kunal Talwar. Private adaptive gradient methods for convex optimization. In *International Conference on Machine Learning*, 2021.

P. Auer, M. Herbster, and M. Warmuth. Exponentially many local minima for single neurons. In *Advances in Neural Information Processing Systems*, 1996.

Haim Avron, Michael Kapralov, Cameron Musco, Christopher Musco, Ameya Velingker, and Amir Zandieh. Random fourier features for kernel ridge regression: Approximation bounds and statistical guarantees. In *International Conference on Machine Learning*, 2017.

Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. In *Advances in Neural Information Processing Systems*, 2022.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*, 2014.

Borja Balle, Giovanni Cherubin, and Jamie Hayes. Reconstructing training data with informed adversaries. In *2022 IEEE Symposium on Security and Privacy (SP)*, 2022.

Rachit Bansal, Danish Pruthi, and Yonatan Belinkov. Measures of information reflect memorization patterns. In *Advances in Neural Information Processing Systems*, 2022.

Rina Foygel Barber and John C Duchi. Privacy and statistical risk: Formalisms and minimax bounds. *arXiv preprint arXiv:1412.4451*, 2014.

Brooks Barnes. Disney and Universal Sue A.I. Firm for Copyright Infringement. The New York Times, 2025. Accessed: 2025-09-13.

Peter Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research (JMLR)*, 20(63):1–17, 2019.

Peter Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 2020.

Peter Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: A statistical viewpoint. *Acta Numerica*, 30:87–201, 2021.

Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, 2017.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, 2014.

Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Guha Thakurta. Private stochastic convex optimization with optimal rates. In *Advances in Neural Information Processing Systems*, 2019.

Raef Bassily, Kobbi Nissim, Adam Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. *SIAM Journal on Computing*, 50(3), 2021.

Raef Bassily, Mehryar Mohri, and Ananda Theertha Suresh. Differentially private learning with margin guarantees. In *Advances in Neural Information Processing Systems*, 2022.

Eric B. Baum. On the capabilities of multilayer perceptrons. *Journal of Complexity*, 4, 1988.

Mikhail Belkin. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numerica*, 30:203–248, 2021.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019a.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 2019b.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

Ido Ben-Shaul, Tomer Galanti, and Shai Dekel. Exploring the approximation capabilities of multiplicative neural networks for smooth functions. *Transactions on Machine Learning Research*, 2023.

Lucas Benigni and Sandrine Péché. Eigenvalue distribution of some nonlinear models of random matrices. *Electronic Journal of Probability*, 26:1 – 37, 2021.

Amanda Bertsch, Uri Alon, Graham Neubig, and Matthew R. Gormley. Unlimiformer: Long-range transformers with unlimited length input. In *Advances in Neural Information Processing Systems*, 2023.

Dimitri Bertsekas. *Network optimization: continuous and discrete models*, volume 8. Athena Scientific, 1998.

Rajendra Bhatia. *Positive Definite Matrices*. Princeton University Press, 2007. ISBN 9780691129181.

Alberto Bietti, Vivien Cabannes, Diane Bouchacourt, Herve Jegou, and Leon Bottou. Birth of a transformer: A memory viewpoint. In *Advances in Neural Information Processing Systems*, 2023.

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Machine Learning and Knowledge Discovery in Databases*, pages 387–402, 2013.

Simone Bombari and Marco Mondelli. How spurious features are memorized: Precise analysis for random and NTK features. In *International Conference on Machine Learning*, 2024a.

Simone Bombari and Marco Mondelli. Towards understanding the word sensitivity of attention layers: A study via random features. In *International Conference on Machine Learning*, 2024b.

Simone Bombari and Marco Mondelli. Privacy for free in the overparameterized regime. *Proceedings of the National Academy of Sciences*, 122(15):e2423072122, 2025a.

Simone Bombari and Marco Mondelli. Spurious correlations in high dimensional regression: The roles of regularization, simplicity bias and over-parameterization. In *International Conference on Machine Learning*, 2025b.

Simone Bombari, Alessandro Achille, Zijian Wang, Yu-Xiang Wang, Yusheng Xie, Kunwar Yashraj Singh, Srikar Appalaraju, Vijay Mahadevan, and Stefano Soatto. Towards differential relational privacy and its use in question answering. *arXiv preprint arXiv:2203.16701*, 2022a.

Simone Bombari, Mohammad Hossein Amani, and Marco Mondelli. Memorization and optimization in deep neural networks with minimum over-parameterization. In *Advances in Neural Information Processing Systems*, 2022b.

Simone Bombari, Shayan Kiyani, and Marco Mondelli. Beyond the universal law of robustness: Sharper laws for random features and neural tangent kernels. In *International Conference on Machine Learning*, 2023.

Simone Bombari, Jialei Luo, Inbar Seroussi, and Marco Mondelli. High-dimensional private linear regression with optimal rates. *arXiv preprint arXiv:2505.16329*, 2026.

David Bosch, Ashkan Panahi, and Babak Hassibi. Precise asymptotic analysis of deep random feature models. In *Annual Conference on Learning Theory (COLT)*, 2023.

Olivier Bousquet and André Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024a.

Gavin Brown, Mark Bun, Vitaly Feldman, Adam Smith, and Kunal Talwar. When is memorization of irrelevant training data necessary for high-accuracy learning? In *Proceedings of the 53rd annual ACM SIGACT symposium on theory of computing*, pages 123–132, 2021.

Gavin R Brown, Krishnamurthy Dj Dvijotham, Georgina Evans, Daogao Liu, Adam Smith, and Abhradeep Guha Thakurta. Private gradient descent for linear regression: Tighter error bounds and instance-specific uncertainty estimation. In *International Conference on Machine Learning*, 2024b.

Tom Brown et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020.

Zhiqi Bu, Hua Wang, Zongyu Dai, and Qi Long. On the convergence and calibration of deep learning with differential privacy. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Sebastien Bubeck and Mark Sellke. A universal law of robustness via isoperimetry. In *Advances in Neural Information Processing Systems*, 2021.

Sébastien Bubeck, Ronen Eldan, Yin Tat Lee, and Dan Mikulincer. Network size and weights size for memorization with two-layers neural networks. In *Advances in Neural Information Processing Systems*, 2020.

Sébastien Bubeck, Yuanzhi Li, and Dheeraj M Nagaraj. A law of robustness for two-layers neural networks. In *Conference on Learning Theory (COLT)*, pages 804–820, 2021.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Sam Buchanan, Dar Gilboa, and John Wright. Deep networks and the multiple manifold problem. In *International Conference on Learning Representations*, 2021.

Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of cryptography conference*, pages 635–658. Springer, 2016.

Mark Bun, Jonathan Ullman, and Salil Vadhan. Fingerprinting codes and the price of approximate differential privacy. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 1–10, 2014.

Gon Buzaglo, Niv Haim, Gilad Yehudai, Gal Vardi, Yakir Oz, Yaniv Nikankin, and Michal Irani. Deconstructing data reconstruction: Multiclass, weight decay and general losses. In *Advances in Neural Information Processing Systems*, 2023.

Vivien Cabannes, Elvis Dohmatob, and Alberto Bietti. Scaling laws for associative memories. In *International Conference on Learning Representations*, 2024.

T. Tony Cai, Yichen Wang, and Linjun Zhang. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *The Annals of Statistics*, 49 (5):2825 – 2850, 2021.

T. Tony Cai, Yichen Wang, and Linjun Zhang. Score attack: A lower bound technique for optimal differentially private learning. *arXiv preprint arXiv:2303.07152*, 2025.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational mathematics*, 7(3):331–368, 2007.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *USENIX Conference on Security Symposium*, 2019.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, Dawn Xiaodong Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *USENIX Conference on Security Symposium*, 2021.

Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *USENIX Security Symposium*, 2023a.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In *International Conference on Learning Representations*, 2023b.

Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems*, 2019.

Michael Celentano, Chen Cheng, and Andrea Montanari. The high-dimensional asymptotics of first order methods with random data. *arXiv preprint arXiv:2112.07572*, 2021.

C. Chang, G. Adam, and A. Goldenberg. Towards robust classification model by counterfactual and invariant data generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021a.

Xiangyu Chang, Yingcong Li, Samet Oymak, and Christos Thrampoulidis. Provable benefits of overparameterization in model compression: From double descent to pruning neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6974–6983, 2021b.

Kamalika Chaudhuri and Claire Monteleoni. Privacy-preserving logistic regression. In *Advances in Neural Information Processing Systems*, 2008.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(29):1069–1109, 2011.

Xiangyi Chen, Steven Z. Wu, and Mingyi Hong. Understanding gradient clipping in private sgd: A geometric perspective. In *Advances in Neural Information Processing Systems*, 2020a.

Zixiang Chen, Yuan Cao, Difan Zou, and Quanquan Gu. How much over-parameterization is sufficient to learn deep relu networks? In *International Conference on Learning Representations*, 2020b.

Chen Cheng and Andrea Montanari. Dimension free ridge regression. *The Annals of Statistics*, 52(6):2879 – 2912, 2024.

Yuri Chervonyi, Trieu H. Trinh, Miroslav Olšák, Xiaomeng Yang, Hoang H. Nguyen, Marcelo Menegali, Junehyuk Jung, Junsu Kim, Vikas Verma, Quoc V. Le, and Thang Luong. Gold-medalist performance in solving olympiad geometry with alphageometry2. *Journal of Machine Learning Research*, 26(241):1–39, 2025.

Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems*, 2018.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, 2019.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30, 2017.

Aloni Cohen. Blameless users in a clean room: Defining copyright protection for generative models. In *Advances in Neural Information Processing Systems*, 2025.

Elizabeth Collins-Woodfin and Inbar Seroussi. Exact dynamics of multi-class stochastic gradient descent. *arXiv preprint arXiv:2510.14074*, 2025.

Elizabeth Collins-Woodfin, Courtney Paquette, Elliot Paquette, and Inbar Seroussi. Hitting the high-dimensional notes: an ode for sgd learning dynamics on glms and multi-index models. *Information and Inference: A Journal of the IMA*, 13(4):iaae028, 2024a. ISSN 2049-8772.

Elizabeth Collins-Woodfin, Inbar Seroussi, Begoña García Malaxechebarría, Andrew W Mackenzie, Elliot Paquette, and Courtney Paquette. The high line: Exact risk and learning rate curves of stochastic adaptive learning rate algorithms. In *Advances in Neural Information Processing Systems*, 2024b.

A Feder Cooper and James Grimmelmann. The files are in the computer: On copyright, memorization, and generative ai. *Cornell Legal Studies Research Paper Forthcoming, Chicago-Kent Law Review, Forthcoming*, 2024.

A Feder Cooper, Aaron Gokaslan, Amy B Cyphert, Christopher De Sa, Mark Lemley, Daniel E Ho, and Percy Liang. Extracting memorized pieces of (copyrighted) books from open-weight language models. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*, 2025.

Thomas M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, EC-14(3): 326–334, 1965.

Rachel Cummings, Varun Gupta, Dhamma Kimpara, and Jamie Morgenstern. On the compatibility of privacy and fairness. In *Adjunct publication of the 27th conference on user modeling, adaptation and personalization*, pages 309–315, 2019.

Alexandru Damian, Jason Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In *Conference on Learning Theory*, pages 5413–5452. PMLR, 2022.

Amit Daniely. Memorizing gaussians with no over-parameterizaion via gradient decent on neural networks. *arXiv preprint arXiv:2003.12895*, 2020a.

Amit Daniely. Neural networks learning and memorization with (almost) no over-parameterization. In *Advances in Neural Information Processing Systems*, 2020b.

Rudrajit Das, Satyen Kale, Zheng Xu, Tong Zhang, and Sujay Sanghavi. Beyond uniform lipschitz condition in differentially private optimization. In *International Conference on Machine Learning*, 2023.

Soham De, Leonard Berrada, Jamie Hayes, Samuel L. Smith, and Borja Balle. Unlocking high-accuracy differentially private image classification through scale. *arXiv preprint arXiv:2204.13650*, 2022.

Leonardo Defilippis, Bruno Loureiro, and Theodor Misiakiewicz. Dimension-free deterministic equivalents and scaling laws for random feature regression. In *Advances in Neural Information Processing Systems*, 2024.

Zeyu Deng, Abla Kammoun, and Christos Thrampoulidis. A model of double descent for high-dimensional binary linear classification. *Information and Inference: A Journal of the IMA*, 11(2):435–495, 2022.

Damien Desfontaines. A list of real-world uses of differential privacy. Personal blog, 10 2021. Ted is writing things.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019.

Lee H. Dicker. Ridge regression and asymptotic minimax estimation over spheres of growing dimension. *Bernoulli*, 22(1):1–37, 2016. ISSN 13507265.

Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018.

Edgar Dobriban, Hamed Hassani, David Hong, and Alexander Robey. Provable tradeoffs in adversarially robust classification. *arXiv preprint arXiv:2006.05161*, 2020.

Elvis Dohmatob. Fundamental tradeoffs between memorization and robustness in random features and neural tangent regimes. *arXiv preprint arXiv:2106.02630*, 2022.

Elvis Dohmatob and Alberto Bietti. On the (non-) robustness of two-layer neural networks in different learning regimes. *arXiv preprint arXiv:2203.11864*, 2022.

Konstantin Donhauser, Alexandru Tifrea, Michael Aerni, Reinhard Heckel, and Fanny Yang. Interpolation can hurt robust generalization even when there is no noise. In *Advances in Neural Information Processing Systems*, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Simon S. Du, Jason D. Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, 2019a.

Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019b.

John Duchi, Martin J Wainwright, and Michael I Jordan. Local privacy and minimax bounds: Sharp rates for probability estimation. In *Advances in Neural Information Processing Systems*, 2013.

John C Duchi and Feng Ruan. The right complexity measure in locally private estimation: It is not the fisher information. *The Annals of Statistics*, 52(1):1–51, 2024.

John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association*, 113(521): 182–201, 2018.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography, Third Theory of Cryptography Conference, TCC 2006*, volume 3876 of *Lecture Notes in Computer Science*, pages 265–284. Springer, 2006.

Cynthia Dwork, Adam Smith, Thomas Steinke, Jonathan Ullman, and Salil Vadhan. Robust traceability from trace amounts. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 650–669. IEEE, 2015.

Cynthia Dwork, Pranay Tankala, and Linjun Zhang. Differentially private learning beyond the classical dimensionality regime. *arXiv preprint arXiv:2411.13682*, 2025.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. HotFlip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018.

Benjamin L Edelman, Surbhi Goel, Sham Kakade, and Cyril Zhang. Inductive biases and variable creation in self-attention mechanisms. In *International Conference on Machine Learning*, 2022.

André Elisseeff and Massimiliano Pontil. Leave-one-out error and stability of learning algorithms with applications stability of randomized learning algorithms source. *International Journal of Systems Science (IJSySc)*, 6, 2002.

Niva Elkin-Koren, Uri Hacohen, Roi Livni, and Shay Moran. Can copyright be reduced to privacy? In *Symposium on Foundations of Responsible Computing*, 2024.

Zhou Fan and Zhichao Wang. Spectra of the conjugate kernel and neural tangent kernel for linear-width neural networks. In *Advances in Neural Information Processing Systems*, 2020.

Vitaly Feldman. Does learning require memorization? A short tale about a long tail. In *ACM Symposium on Theory of Computing (STOC)*, pages 954–959, 2020.

Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. In *Advances in Neural Information Processing Systems*, 2020.

Vitaly Feldman, Ilya Mironov, Kunal Talwar, and Abhradeep Thakurta. Privacy Amplification by Iteration. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 521–532, 2018.

Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: optimal rates in linear time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, 2020.

Vitaly Feldman, Guy Kornowski, and Xin Lyu. Trade-offs in data memorization via strong data processing inequalities. *arXiv preprint arXiv:2506.01855*, 2025.

Damien Ferbach, Katie Everett, Gauthier Gidel, Elliot Paquette, and Courtney Paquette. Dimension-adapted momentum outscales SGD. In *Advances in Neural Information Processing Systems*, 2025.

Ferdinando Fioretto, Cuong Tran, Pascal Van Hentenryck, and Keyu Zhu. Differential privacy and fairness in decisions and learning tasks: A survey. *arXiv preprint arXiv:2202.08187*, 2022.

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1322–1333, 2015.

Hengyu Fu, Tianyu Guo, Yu Bai, and Song Mei. What can a single attention layer learn? a study through the random features lens. In *Advances in Neural Information Processing Systems*, 2023.

C.W. Gardiner. *Handbook of Stochastic Methods for Physics, Chemistry, and the Natural Sciences*. Proceedings in Life Sciences. Springer-Verlag, 1985. ISBN 9783540113577.

Rong Ge, Runzhe Wang, and Haoyu Zhao. Mildly overparametrized neural nets can memorize training data efficiently. *arXiv preprint arXiv:1909.11837*, 2019.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Cedric Gerbelot, Emanuele Troiani, Francesca Mignacco, Florent Krzakala, and Lenka Zdeborova. Rigorous dynamical mean-field theory for stochastic gradient descent methods. *SIAM Journal on Mathematics of Data Science*, 6(2):400–427, 2024.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. When do neural networks outperform kernel methods? In *Advances in Neural Information Processing Systems*, 2020.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *The Annals of Statistics*, 49(2):1029–1054, 2021.

Bishwamittra Ghosh, Debabrota Basu, and Kuldeep S. Meel. "How biased are your features?": Computing fairness influence functions with global sensitivity analysis. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2023.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *International Conference on Machine Learning*, 2010.

Aditya Golatkar, Alessandro Achille, Yu-Xiang Wang, Aaron Roth, Michael Kearns, and Stefano Soatto. Mixed differential privacy in computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Sebastian Goldt, Marc Mézard, Florent Krzakala, and Lenka Zdeborová. Modeling the influence of data structure on learning in neural networks: The hidden manifold model. *Physical Review X*, 10(4):041044, 2020.

Sebastian Goldt, Bruno Loureiro, Galen Reeves, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. The gaussian equivalence of generative models for learning with shallow neural networks. In *Mathematical and Scientific Machine Learning*, pages 426–471. PMLR, 2022.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, 2017.

Chuan Guo, Alexandre Sablayrolles, Hervé Jégou, and Douwe Kiela. Gradient-based adversarial attacks against text transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

Niv Haim, Gal Vardi, Gilad Yehudai, Ohad Shamir, and Michal Irani. Reconstructing training data from trained neural networks. In *Advances in Neural Information Processing Systems*, 2022.

Rob Hall, Alessandro Rinaldo, and Larry Wasserman. Random differential privacy. *Journal of Privacy and Confidentiality*, 4(2), 2013.

Dongyoon Han, Jiwhan Kim, and Junmo Kim. Deep pyramidal residual networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5927–5935, 2017.

Qiyang Han. Entrywise dynamics and universality of general first order methods. *arXiv preprint arXiv:2406.19061*, 2024.

Qiyang Han and Xiaocong Xu. The distribution of ridgeless least squares interpolators. *arXiv preprint arXiv:2307.02044*, 2023.

Boris Hanin. Which neural net architectures give rise to exploding and vanishing gradients? In *Advances in Neural Information Processing Systems*, 2018.

Boris Hanin. Universal function approximation by deep neural nets with bounded width and relu activations. *Mathematics*, 7(10), 2019.

Moritz Hardt, Benjamin Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, 2016.

Hamed Hassani and Adel Javanmard. The curse of overparametrization in adversarial training: Precise analysis of robust generalization for random features regression. *The Annals of Statistics*, 52(2):441–465, 2024.

Babak Hassibi and David Stork. Second order derivatives for network pruning: Optimal brain surgeon. In *Advances in Neural Information Processing Systems*, 1992.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022.

Michael B. Hawes. Implementing Differential Privacy: Seven Lessons From the 2020 United States Census. *Harvard Data Science Review*, 2(2), apr 30 2020. Harvard Data Science Review.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Conference on Computer Vision and Pattern Recognition*, 2016.

Katherine Hermann and Andrew Lampinen. What shapes feature representations? Exploring datasets, architectures, and training. In *Advances in Neural Information Processing Systems*, 2020.

Katherine Hermann, Hossein Mobahi, Thomas FEL, and Michael Curtis Mozer. On the foundations of shortcut learning. In *International Conference on Learning Representations*, 2024.

Magnus R Hestenes, Eduard Stiefel, et al. Methods of conjugate gradients for solving linear systems. *Journal of research of the National Bureau of Standards*, 49(6):409–436, 1952.

David C. Hoaglin and Roy E. Welsch. The hat matrix in regression and anova. *The American Statistician*, 32(1):17–22, 1978.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack William Rae, and Laurent Sifre. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems*, 2022.

Hong Hu and Yue M Lu. Universality laws for high-dimensional learning with random features. *IEEE Transactions on Information Theory*, 69(3):1932–1964, 2022.

Hong Hu, Yue M. Lu, and Theodor Misiakiewicz. Asymptotics of random feature regression beyond the linear scaling regime. *arXiv preprint arXiv:2403.08160*, 2024.

Wei Hu, Lechao Xiao, Ben Adlam, and Jeffrey Pennington. The surprising simplicity of the early-time learning dynamics of neural networks. In *Advances in Neural Information Processing Systems*, 2020.

Baihe Huang, Xiaoxiao Li, Zhao Song, and Xin Yang. Fl-ntk: A neural tangent kernel-based framework for federated learning analysis. In *International Conference on Machine Learning*, 2021a.

Hanxun Huang, Yisen Wang, Sarah Erfani, Quanquan Gu, James Bailey, and Xingjun Ma. Exploring architectural ingredients of adversarially robust deep neural networks. In *Advances in Neural Information Processing Systems*, 2021b.

Peter J. Huber. Robust Regression: Asymptotics, Conjectures and Monte Carlo. *The Annals of Statistics*, 1(5):799 – 821, 1973.

Leonardo Iurada, Simone Bombari, Tatiana Tommasi, and Marco Mondelli. A law of data reconstruction for random features (and beyond). In *International Conference on Learning Representations*, 2026.

Pavel Izmailov, Polina Kirichenko, Nate Gruver, and Andrew Gordon Wilson. On feature learning in the presence of spurious correlations. In *Advances in Neural Information Processing Systems*, 2022.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, 2018.

Prateek Jain and Abhradeep Guha Thakurta. (near) dimension independent risk bounds for differentially private learning. In *International Conference on Machine Learning*, 2014.

Adel Javanmard and Mahdi Soltanolkotabi. Precise statistical analysis of classification accuracies for adversarial training. *The Annals of Statistics*, 50(4):2127–2156, 2022.

Adel Javanmard, Marco Mondelli, and Andrea Montanari. Analysis of a two-layer neural network via displacement convexity. *Annals of Statistics*, 2020a.

Adel Javanmard, Mahdi Soltanolkotabi, and Hamed Hassani. Precise tradeoffs in adversarial training for linear regression. In *Conference on Learning Theory (COLT)*, 2020b.

Samy Jelassi, Michael Eli Sander, and Yuanzhi Li. Vision transformers provably learn spatial structure. In *Advances in Neural Information Processing Systems*, 2022.

Ziwei Ji and Matus Telgarsky. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks. In *International Conference on Learning Representations*, 2019.

Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. In *Advances in Neural Information Processing Systems*, 2020.

Charles R. Johnson. *Matrix Theory and Applications*. American Mathematical Society, 1990.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 2021.

Peter Kairouz, Brendan Mcmahan, Shuang Song, Om Thakkar, Abhradeep Thakurta, and Zheng Xu. Practical and private (deep) learning without sampling or shuffling. In *International Conference on Machine Learning*, 2021.

Dimitris Kalimeris, Gal Kaplun, Preetum Nakkiran, Benjamin Edelman, Tristan Yang, Boaz Barak, and Haofeng Zhang. Sgd on neural networks learns functions of increasing complexity. In *Advances in Neural Information Processing Systems*, 2019.

Gautam Kamath. Private ML and Stats: Modern ML. Lecture notes, 2020. CS 860: Algorithms for Private Data Analysis, Fall 2020, Lecture 14, University of Waterloo.

Gautam Kamath, Jerry Li, Vikrant Singhal, and Jonathan Ullman. Privately learning high-dimensional distributions. In *Proceedings of the Thirty-Second Conference on Learning Theory*. PMLR, 2019.

Gautam Kamath, Vikrant Singhal, and Jonathan Ullman. Private mean estimation of heavy-tailed distributions. In *Proceedings of Thirty Third Conference on Learning Theory*, 2020.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Daniel Kifer, Adam Smith, and Abhradeep Thakurta. Private convex empirical risk minimization and high-dimensional regression. In *Proceedings of the 25th Annual Conference on Learning Theory*, 2012.

Hyunjik Kim, George Papamakarios, and Andriy Mnih. The lipschitz constant of self-attention. In *International Conference on Machine Learning*, 2021.

Yoon Kim, Carl Denton, Luong Hoang, and Alexander M. Rush. Structured attention networks. In *International Conference on Learning Representations*, 2017.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *International Conference on Learning Representations*, 2023.

P.E. Kloeden and E. Platen. *Numerical Solution of Stochastic Differential Equations*. Stochastic Modelling and Applied Probability. Springer Berlin Heidelberg, 2011. ISBN 9783540540625.

Anastasia Koloskova, Youssef Allouah, Animesh Jha, Rachid Guerraoui, and Sanmi Koyejo. Certified unlearning for neural networks. In *International Conference on Machine Learning*, 2025.

Vladimir Koltchinskii and Dmitriy Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*, 30(1):1–50, 2002.

Adam Kowalczyk. Counting function theorem for multi-layer networks. In *Advances in Neural Information Processing Systems*, 1994.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Master's thesis, University of Toronto, ON, Canada, 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 2017.

Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *International Conference on Learning Representations*, 2017.

Alexey Kurakin, Shuang Song, Steve Chien, Roxana Geambasu, Andreas Terzis, and Abhradeep Thakurta. Toward training at imagenet scale with differential privacy. *arXiv preprint arXiv:2201.12328*, 2022.

Yann Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Yann LeCun, Léon Bottou, Genevieve Orr, and Klaus-Robert Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–48. Springer, 2012.

Donghwan Lee, Behrad Moniri, Xinmeng Huang, Edgar Dobriban, and Hamed Hassani. Demystifying disagreement-on-the-line in high dimensions. In *International Conference on Machine Learning*, 2023.

Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in Neural Information Processing Systems*, 2019.

Klas Leino and Matt Fredrikson. Stolen memories: leveraging model memorization for calibrated white-box membership inference. In *Proceedings of the 29th USENIX Conference on Security Symposium*, 2020.

Hongkang Li, Meng Wang, Sijia Liu, and Pin-Yu Chen. A theoretical understanding of shallow vision transformers: Learning, generalization, and sample complexity. In *International Conference on Learning Representations*, 2023.

Xuechen Li, Daogao Liu, Tatsunori Hashimoto, Huseyin A Inan, Janardhan Kulkarni, YinTat Lee, and Abhradeep Guha Thakurta. When does differentially private learning not suffer in high dimensions? In *Advances in Neural Information Processing Systems*, 2022a.

Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. In *International Conference on Learning Representations*, 2022b.

Zhenyu Liao and Romain Couillet. On the spectrum of random features maps of high dimensional data. In *International Conference on Machine Learning*, 2018.

Zhenyu Liao, Romain Couillet, and Michael W. Mahoney. A random matrix analysis of random Fourier features: beyond the Gaussian kernel, a precise phase transition, and the corresponding double descent. In *Advances in Neural Information Processing Systems*, 2020.

Licong Lin, Jingfeng Wu, Sham M. Kakade, Peter L. Bartlett, and Jason D. Lee. Scaling laws in linear regression: Compute, parameters, and data. In *Advances in Neural Information Processing Systems*, 2024.

Licong Lin, Jingfeng Wu, and Peter L Bartlett. Improved scaling laws in linear regression via data reuse. *arXiv preprint arXiv:2506.08415*, 2025a.

Shurong Lin, Eric D Kolaczyk, Adam Smith, and Elliot Paquette. High-dimensional privacy-utility dynamics of noisy stochastic gradient descent on least squares. *arXiv preprint arXiv:2510.16687*, 2025b.

Chaoyue Liu, Libin Zhu, and Misha Belkin. On the linearity of large non-linear models: when and why the tangent kernel is constant. In *Advances in Neural Information Processing Systems*, 2020.

Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, 2021.

Sheng Liu, Xu Zhang, Nitesh Sekhar, Yue Wu, Prateek Singhal, and Carlos Fernandez-Granda. Avoiding spurious correlations via logit correction. In *International Conference on Learning Representations*, 2023a.

Xiyang Liu, Weihao Kong, and Sewoong Oh. Differential privacy and robust statistics in high dimensions. In *Proceedings of Thirty Fifth Conference on Learning Theory*, 2022.

Xiyang Liu, Prateek Jain, Weihao Kong, Sewoong Oh, and Arun Suggala. Label robust and differentially private linear regression: Computational and statistical efficiency. In *Advances in Neural Information Processing Systems*, 2023b.

Noel Loo, Ramin Hasani, Alexander Amini, and Daniela Rus. Evolution of neural tangent kernels under benign and adversarial training. In *Advances in Neural Information Processing Systems*, 2022.

Noel Loo, Ramin Hasani, Mathias Lechner, Alexander Amini, and Daniela Rus. Understanding reconstruction attacks with the neural tangent kernel and dataset distillation. In *International Conference on Learning Representations*, 2024.

Cosme Louart, Zhenyu Liao, and Romain Couillet. A random matrix approach to neural networks. *The Annals of Applied Probability*, 28(2):1190–1248, 2018.

Bruno Loureiro, Cedric Gerbelot, Hugo Cui, Sebastian Goldt, Florent Krzakala, Marc Mezard, and Lenka Zdeborová. Learning curves of generic features maps for realistic datasets with a teacher-student model. In *Advances in Neural Information Processing Systems*, 2021.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2020.

Yi-An Ma, Teodor Vanislavov Marinov, and Tong Zhang. Dimension independent generalization of dp-sgd for overparameterized smooth convex optimization. *arXiv preprint arXiv:2206.01836*, 2022.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, 2011.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5188–5196, 2015.

Neil Rohit Mallinar, Austin Zane, Spencer Frei, and Bin Yu. Minimum-norm interpolation under covariate shift. In *International Conference on Machine Learning*, 2024.

Noah Marshall, Ke Liang Xiao, Atish Agarwala, and Elliot Paquette. To clip or not to clip: the dynamics of SGD with gradient clipping in high-dimensions. In *International Conference on Learning Representations*, 2025.

Ryan McKenna, Yangsibo Huang, Amer Sinha, Borja Balle, Zachary Charles, Christopher A. Choquette-Choo, Badih Ghazi, George Kaissis, Ravi Kumar, Ruibo Liu, Da Yu, and Chiyuan Zhang. Scaling laws for differentially private language models. *arXiv preprint arXiv:2501.18914*, 2025.

H. Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. In *International Conference on Learning Representations*, 2018.

Harsh Mehta, Abhradeep Thakurta, Alexey Kurakin, and Ashok Cutkosky. Large scale transfer learning for differentially private image classification. *arXiv preprint arXiv:2205.02973*, 2022.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 2018.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Conference on Learning Theory*, pages 2388–2464. PMLR, 2019.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Generalization error of random feature and kernel methods: hypercontractivity and kernel matrix concentration. *Applied and Computational Harmonic Analysis*, 2021.

S. Menozzi, A. Pesce, and X. Zhang. Density and gradient estimates for non degenerate brownian sdes with unbounded measurable drift. *Journal of Differential Equations*, 272: 330–369, 2021. ISSN 0022-0396.

Jason Milionis, Alkis Kalavasis, Dimitris Fotakis, and Stratis Ioannidis. Differentially private regression with unbounded covariates. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, 2022.

Yifei Min, Lin Chen, and Amin Karbasi. The curious case of adversarially robust models: More data can help, double descend, or hurt generalization. In *Uncertainty in Artificial Intelligence (UAI)*, 2021.

Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275, 2017.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

Mazda Moayeri, Sahil Singla, and Soheil Feizi. Hard imagenet: Segmentations for objects with strong spurious cues. In *Advances in Neural Information Processing Systems*, 2022.

Behrad Moniri, Donghwan Lee, Hamed Hassani, and Edgar Dobriban. A theory of non-linear feature learning with one gradient step in two-layer neural networks. *arXiv preprint arXiv:2310.07891*, 2023.

Andrea Montanari and Basil N Saeed. Universality of empirical risk minimization. In *Conference on Learning Theory*, pages 4310–4312. PMLR, 2022.

Andrea Montanari and Yiqiao Zhong. The interpolation phase transition in neural networks: Memorization and generalization under lazy training. *arXiv preprint arXiv:2007.12826*, 2020.

Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. The generalization error of max-margin linear classifiers: Benign overfitting and high dimensional asymptotics in the overparametrized regime. *The Annals of Statistics*, 53(2):822–853, 2025.

Depen Morwani, jatin batra, Prateek Jain, and Praneeth Netrapalli. Simplicity bias in 1-hidden layer neural networks. In *Advances in Neural Information Processing Systems*, 2023.

Jaouad Mourtada. Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices. *Annals of Statistics*, 50(4), 2022.

Sayan Mukherjee, Partha Niyogi, Tomaso Poggio, and Ryan Rifkin. Learning theory: Stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Adv. Comput. Math.*, 25:161–193, 2006.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. In *International Conference on Learning Representations*, 2020.

Milad Nasr, Javier Rando, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Florian Tramèr, and Katherine Lee. Scalable extraction of training data from aligned, production language models. In *International Conference on Learning Representations*, 2025a.

Milad Nasr, Thomas Steinke, Borja Balle, Christopher A. Choquette-Choo, Arun Ganesh, Matthew Jagielski, Jamie Hayes, Abhradeep Guha Thakurta, Adam Smith, and Andreas Terzis. The last iterate advantage: Empirical auditing and principled heuristic analysis of differentially private SGD. In *International Conference on Learning Representations*, 2025b.

Quynh Nguyen. On the proof of global convergence of gradient descent for deep relu networks with linear widths. In *International Conference on Machine Learning*, 2021.

Quynh Nguyen and Matthias Hein. The loss surface of deep and wide neural networks. In *International Conference on Machine Learning*, 2017.

Quynh Nguyen and Matthias Hein. Optimization landscape and expressivity of deep CNNs. In *International Conference on Machine Learning*, 2018.

Quynh Nguyen and Marco Mondelli. Global convergence of deep networks with one wide layer followed by pyramidal topology. In *Advances in Neural Information Processing Systems*, 2020.

Quynh Nguyen, Pierre Bréchet, and Marco Mondelli. When are solutions connected in deep networks? In *Advances in Neural Information Processing Systems*, 2021a.

Quynh Nguyen, Marco Mondelli, and Guido Montufar. Tight bounds on the smallest eigenvalue of the neural tangent kernel for deep ReLU networks. In *International Conference on Machine Learning*, 2021b.

Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014.

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

OpenAI. Learning to Reason with LLMs. OpenAI, 2024. Accessed: 2026-05-25.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

Samet Oymak and Mahdi Soltanolkotabi. Overparameterized nonlinear learning: Gradient descent takes the shortest path? In *International Conference on Machine Learning*, 2019.

Samet Oymak and Mahdi Soltanolkotabi. Toward moderate overparameterization: Global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 1(1):84–105, 2020.

Samet Oymak, Ankit Singh Rawat, Mahdi Soltanolkotabi, and Christos Thrampoulidis. On the role of attention in prompt-tuning. In *International Conference on Machine Learning*, 2023.

Yakir Oz, Gilad Yehudai, Gal Vardi, Itai Antebi, Michal Irani, and Niv Haim. Reconstructing training data from real world models trained with transfer learning. *arXiv preprint arXiv:2407.15845*, 2024.

Parthe Pandit, Zhichao Wang, and Yizhe Zhu. Universality of kernel random matrices and kernel regression in the quadratic regime. *arXiv preprint arXiv:2408.01062*, 2024.

Nicolas Papernot, Abhradeep Thakurta, Shuang Song, Steve Chien, and Úlfar Erlingsson. Tempered sigmoid activations for deep learning with differential privacy. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

Courtney Paquette, Elliot Paquette, Ben Adlam, and Jeffrey Pennington. Implicit regularization or implicit conditioning? exact risk trajectories of sgd in high dimensions. In *Advances in Neural Information Processing Systems*, 2022.

Courtney Paquette, Elliot Paquette, Ben Adlam, and Jeffrey Pennington. Homogenization of sgd in high-dimensions: exact dynamics and generalization properties. *Mathematical Programming*, 2024a.

Elliot Paquette, Courtney Paquette, Lechao Xiao, and Jeffrey Pennington. 4+3 phases of compute-optimal neural scaling laws. In *Advances in Neural Information Processing Systems*, 2024b.

Sandrine Péché. A note on the Pennington-Worah distribution. *Electronic Communications in Probability*, 24:1–7, 2019.

Jeffrey Pennington and Yasaman Bahri. Geometry of neural network loss surfaces via random matrix theory. In *International Conference on Machine Learning*, 2017.

Jeffrey Pennington and Pratik Worah. Nonlinear random matrix theory for deep learning. In *Advances in Neural Information Processing Systems*, 2017.

Jeffrey Pennington and Pratik Worah. The spectrum of the fisher information matrix of a single-hidden-layer neural network. In *Advances in Neural Information Processing Systems*, 2018.

Jeffrey Pennington, Samuel Schoenholz, and Surya Ganguli. The emergence of spectral universality in deep networks. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018.

Mohammad Pezeshki, Sékou-Oumar Kaba, Yoshua Bengio, Aaron Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. In *Advances in Neural Information Processing Systems*, 2021.

Vanessa Piccolo and Dominik Schröder. Analysis of one-hidden-layer neural networks via the resolvent method. In *Advances in Neural Information Processing Systems*, 2021.

Venkatadheeraj Pichapati, Ananda Theertha Suresh, Felix X. Yu, Sashank J. Reddi, and Sanjiv Kumar. Adaclip: Adaptive clipping for private sgd. *arXiv preprint arXiv:1908.07643*, 2019.

Yaniv Plan and Roman Vershynin. Random matrices acting on sets: Independent columns. *arXiv preprint arXiv:2502.16827*, 2025.

Gregory Plumb, Marco Tulio Ribeiro, and Ameet Talwalkar. Finding and fixing spurious patterns with explanations. *Transactions on Machine Learning Research*, 2022.

Audrey Pope. NYT v. OpenAI: The Times's About-Face. Harvard Law Review Blog, 2024. Accessed: 2025-09-13.

Bernd Prach and Christoph H. Lampert. Almost-orthogonal layers for efficient general-purpose lipschitz networks. In *European Conference on Computer Vision (ECCV)*, 2022.

GuanWen Qiu, Da Kuang, and Surbhi Goel. Complexity matters: Dynamics of feature learning in the presence of spurious correlations. In *NeurIPS Workshop on Mathematics of Modern Machine Learning*, 2023.

Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. In *International Conference on Learning Representations*, 2018.

Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John C Duchi, and Percy Liang. Adversarial training can hurt generalization. *arXiv preprint arXiv:1906.06032*, 2019.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International Conference on Machine Learning*, 2019.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, 2007.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, 2021.

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Early stopping and non-parametric regression: An optimal data-dependent stopping rule. *Journal of Machine Learning Research*, 15(11):335–366, 2014.

Saksham Rastogi, Pratyush Maini, and Danish Pruthi. STAMP your content: Proving dataset membership via watermarked rephrasings. In *International Conference on Machine Learning*, 2025.

Yunwei Ren, Eshaan Nichani, Denny Wu, and Jason D. Lee. Emergence and scaling laws in SGD learning of shallow neural networks. In *Advances in Neural Information Processing Systems*, 2025.

Daniel Revuz and Marc Yor. *Continuous Martingales and Brownian Motion*. Grundlehren der mathematischen Wissenschaften. Springer Berlin, Heidelberg, 2010. ISBN 978-3-540-64325-8.

Philippe Rigollet and Jan-Christian Hütter. High-dimensional statistics. *arXiv preprint arXiv:2310.19244*, 2023.

Angelika Rohde and Lukas Steinberger. Geometrizing rates of convergence under local differential privacy constraints. *The Annals of Statistics*, 48(5):2646–2670, 2020.

Valentin Roth. On mcdiarmid's inequality under dependence via approximate tensorization of entropy. *arXiv preprint arXiv:2606.12720*, 2026.

Grant M. Rotskoff and Eric Vanden-Eijnden. Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error. In *Advances in Neural Information Processing Systems*, 2018.

Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems*, 2017.

Alfred Rényi. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 547–561, 1961.

Itay Safran and Ohad Shamir. Spurious local minima are common in two-layer ReLU neural networks. In *International Conference on Machine Learning*, 2018.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020a.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, 2020b.

Akito Sakurai. nh-1 networks store no less n* h+ 1 examples, but sometimes no more. In *IJCNN International Joint Conference on Neural Networks*, 1992.

Holger Sambale. Some notes on concentration for α-subexponential random variables. *arXiv preprint arXiv:2002.10761*, 2020.

Dominik Schröder, Hugo Cui, Daniil Dmitriev, and Bruno Loureiro. Deterministic equivalent and error universality of deep random features learning. In *International Conference on Machine Learning*, 2023.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

J. Schur. Bemerkungen zur theorie der beschränkten bilinearformen mit unendlich vielen veränderlichen. *Journal für die reine und angewandte Mathematik*, 140:1–28, 1911.

Avi Schwarzschild, Zhili Feng, Pratyush Maini, Zachary Chase Lipton, and J Zico Kolter. Rethinking LLM memorization through the lens of adversarial compression. In *Advances in Neural Information Processing Systems*, 2024.

Mohamed El Amine Seddik, Cosme Louart, Mohamed Tamaazousti, and Romain Couillet. Random matrix theory proves that deep learning representations of GAN-data behave as gaussian mixtures. In *International Conference on Machine Learning*, 2020.

Seonguk Seo, Joon-Young Lee, and Bohyung Han. Information-theoretic bias reduction via causal view of spurious correlation. In *AAAI Conference on Artificial Intelligence*, 2022.

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. In *Advances in Neural Information Processing Systems*, 2020.

Shai Shalev-Shwartz, Nathan Srebro, and Karthik Sridharan. Stochastic convex optimization. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT)*, 2009.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

A. E. Shevchenko, Vyacheslav Kungurtsev, and Marco Mondelli. Mean-field analysis of piecewise linear solutions for wide relu networks. *Journal of Machine Learning Research*, 23: 130:1–130:55, 2021.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

Reza Shokri and Vitaly Shmatikov. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, 2015.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

Sahil Singla and Soheil Feizi. Salient imagenet: How to discover spurious features in deep learning? In *International Conference on Learning Representations*, 2022.

Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A law of large numbers. *SIAM Journal on Applied Mathematics*, 80(2):725–752, 2020.

Guy Smorodinsky, Gal Vardi, and Itay Safran. Provable privacy attacks on trained shallow neural networks. *arXiv preprint arXiv:2410.07632*, 2024.

Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations*, 2025.

Mahdi Soltanolkotabi, Adel Javanmard, and Jason D Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 65(2):742–769, 2018.

Chaehwan Song, Ali Ramezani-Kebrya, Thomas Pethick, Armin Eftekhari, and Volkan Cevher. Subquadratic overparameterization for shallow neural networks. In *Advances in Neural Information Processing Systems*, 2021a.

Shuang Song, Kamalika Chaudhuri, and Anand D. Sarwate. Stochastic gradient descent with differentially private updates. In *2013 IEEE Global Conference on Signal and Information Processing*, 2013.

Shuang Song, Thomas Steinke, Om Thakkar, and Abhradeep Thakurta. Evading the curse of dimensionality in unconstrained private glms. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, 2021b.

Yanke Song, Sohom Bhattacharya, and Pragya Sur. Generalization error of min-norm interpolators in transfer learning. *arXiv preprint arXiv:2406.13944*, 2024.

Zhao Song and Xin Yang. Quadratic suffices for over-parametrization via matrix chernoff bound. *arXiv preprint arXiv:1906.03593*, 2020.

Cory Stephenson, suchismita padhy, Abhinav Ganesh, Yue Hui, Hanlin Tang, and SueYeon Chung. On the geometry of generalization and memorization in deep neural networks. In *International Conference on Learning Representations*, 2021.

Mihailo Stojnic. A framework to characterize performance of lasso algorithms. *arXiv preprint arXiv:1303.7291*, 2013.

Pragya Sur and Emmanuel J Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525, 2019.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

Hossein Taheri, Ramtin Pedarsani, and Christos Thrampoulidis. Asymptotic behavior of adversarial training in binary classification. *arXiv preprint arXiv:2010.13275*, 2020.

Gerald Teschl. *Ordinary differential equations and dynamical systems*. American Mathematical Society, 2012.

Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. Regularized linear regression: A precise analysis of the estimation error. In *Conference on Learning Theory*, pages 1683–1709. PMLR, 2015.

Yuandong Tian, Yiping Wang, Beidi Chen, and Simon Shaolei Du. Scan and snap: Understanding training dynamics and token composition in 1-layer transformer. In *Advances in Neural Information Processing Systems*, 2023.

Rishabh Tiwari and Pradeep Shenoy. Overcoming simplicity bias in deep networks using a feature sieve. In *International Conference on Machine Learning*, 2023.

Hugo Touvron et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Florian Tramer and Dan Boneh. Differentially private learning needs better features (or much more data). In *International Conference on Learning Representations*, 2021.

Florian Tramèr, Gautam Kamath, and Nicholas Carlini. Position: Considerations for differentially private learning with large-scale public pretraining. In *International Conference on Machine Learning*, 2024.

Jacob Trauger and Ambuj Tewari. Sequence length independent norm-based generalization bounds for transformers. *arXiv preprint arXiv:2310.13088*, 2023.

Aleksei Triastcyn and Boi Faltings. Bayesian differential privacy for machine learning. In *International Conference on Machine Learning*, 2020.

Trieu H. Trinh, Yuhuai Wu, Quoc V. Le, Thang Luong, David Silver, et al. Olympiad-level formal mathematical reasoning with reinforcement learning. *Nature*, 2025.

Nilesh Tripuraneni, Ben Adlam, and Jeffrey Pennington. Overparameterization improves robustness to covariate shift in high dimensions. In *Advances in Neural Information Processing Systems*, 2021.

Asher Trockman and J. Zico Kolter. Mimetic initialization of self-attention layers. In *International Conference on Machine Learning*, 2023.

Joel Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, page 389–434, 2012.

Nikolaos Tsilivis and Julia Kempe. What can the neural tangent kernel tell us about adversarial robustness? In *Advances in Neural Information Processing Systems*, 2022.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York, 2009. ISBN 978-0-387-79052-7.

U.S. Copyright Office. Copyright and Artificial Intelligence, Part 3: Generative AI Training (Pre-Publication Version). Report, may 2025. Pre-publication version.

Ramon van Handel. Probability in high dimensions, 2016.

Gal Vardi, Gilad Yehudai, and Ohad Shamir. On the optimal memorization power of reLU neural networks. In *International Conference on Learning Representations*, 2022.

Prateek Varshney, Abhradeep Thakurta, and Prateek Jain. (nearly) optimal private linear regression for sub-gaussian data via adaptive clipping. In *Proceedings of Thirty Fifth Conference on Learning Theory*, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. Counterfactual invariance to spurious correlations in text classification. In *Advances in Neural Information Processing Systems*, 2021.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Yonina C. Eldar and Gitta Kutyniok, editors, *Compressed Sensing: Theory and Applications*, page 210–268. Cambridge University Press, 2012.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. Cambridge university press, 2018.

Roman Vershynin. Memory capacity of neural networks with threshold and rectified linear unit activations. *SIAM Journal on Mathematics of Data Science*, 2(4):1004–1033, 2020.

Jonas von Berg, Adalbert Fono, Massimiliano Datres, Sohir Maskey, and Gitta Kutyniok. The price of robustness: Stable classifiers need overparameterization. In *High-dimensional Learning Dynamics 2025*, 2025.

Nikhil Vyas, Sham M. Kakade, and Boaz Barak. On provable copyright protection for generative models. In *International Conference on Machine Learning*, 2023.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

Chendi Wang, Yuqing Zhu, Weijie J Su, and Yu-Xiang Wang. Neural collapse meets differential privacy: Curious behaviors of noisyGD with near-perfect representation learning. In *International Conference on Machine Learning*, 2024.

Di Wang, Changyou Chen, and Jinhui Xu. Differentially private empirical risk minimization with non-convex loss functions. In *International Conference on Machine Learning*, 2019.

Erchi Wang, Yuqing Zhu, and Yu-Xiang Wang. Adapting to linear separable subsets with large-margin in differentially private learning. *arXiv preprint arXiv:2505.24737*, 2025.

Jun-Kun Wang, Chi-Heng Lin, and Jacob D Abernethy. A modular analysis of provable acceleration via Polyak's momentum: Training a wide ReLU network and a deep linear network. In *International Conference on Machine Learning*, 2021.

Yu-Xiang Wang. Revisiting differentially private linear regression: optimal and adaptive prediction & estimation in unbounded domain. *arXiv preprint arXiv:1803.02596*, 2018.

Yunjuan Wang, Enayat Ullah, Poorya Mianjy, and Raman Arora. Adversarial robustness is at odds with lazy training. In *Advances in Neural Information Processing Systems*, 2022.

Zhichao Wang and Yizhe Zhu. Deformed semicircle law and concentration of nonlinear random matrices for ultra-wide neural networks. *arXiv preprint arXiv:2109.09304*, 2021.

Zhichao Wang and Yizhe Zhu. Overparameterized random feature regression with nearly orthogonal data. In *International Conference on Artificial Intelligence and Statistics*, 2023.

Zihan Wang, Jason Lee, and Qi Lei. Reconstructing training data from model gradient, provably. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022.

Johnny Wei, Ryan Wang, and Robin Jia. Proving membership in llm pretraining data via data watermarks. In *Findings of the Association for Computational Linguistics ACL 2024*, 2024.

Tsui-Wei Weng, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, Dong Su, Yupeng Gao, Cho-Jui Hsieh, and Luca Daniel. Evaluating the robustness of neural networks: An extreme value theory approach. In *International Conference on Learning Representations*, 2018.

Oliver Williams and Frank Mcsherry. Probabilistic inference and differential privacy. In *Advances in Neural Information Processing Systems*, 2010.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3–4):229–256, 1992.

Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning*, 2018.

Boxi Wu, Jinghui Chen, Deng Cai, Xiaofei He, and Quanquan Gu. Do wider neural networks really help adversarial robustness? In *Advances in Neural Information Processing Systems*, 2021.

Diyuan Wu, Lehan Chen, Theodor Misiakiewicz, and Marco Mondelli. Improved scaling laws via weak-to-strong generalization in random feature ridge regression. *arXiv preprint arXiv:2603.05691*, 2026.

Xiaoxia Wu, Simon S. Du, and Rachel Ward. Global convergence of adaptive gradient methods for an over-parameterized neural network. *arXiv preprint arXiv:1902.07111*, 2019.

Yongtao Wu, Fanghui Liu, Grigorios Chrysos, and Volkan Cevher. On the convergence of encoder-only shallow transformers. In *Advances in Neural Information Processing Systems*, 2023.

Kai Yuanqing Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. In *International Conference on Learning Representations*, 2021.

Fan Yang, Hongyang R. Zhang, Sen Wu, Christopher Ré, and Weijie J. Su. Precise high-dimensional asymptotics for quantifying heterogeneous transfers. *arXiv preprint arXiv:2010.11750*, 2023.

Yao-Yuan Yang, Chi-Ning Chou, and Kamalika Chaudhuri. Understanding rare spurious correlations in neural networks. In *ICML Workshop on Spurious Correlations, Invariance and Stability*, 2022.

Wenqian Ye, Guangtao Zheng, Xu Cao, Yunsheng Ma, and Aidong Zhang. Spurious correlations in machine learning: A survey. *arXiv preprint arXiv:2402.12715*, 2024.

Da Yu, Huishuai Zhang, Wei Chen, and Tie-Yan Liu. Do not let privacy overbill utility: Gradient embedding perturbation for private learning. In *International Conference on Learning Representations*, 2021a.

Da Yu, Huishuai Zhang, Wei Chen, Jian Yin, and Tie-Yan Liu. Large scale private learning via low-rank reparametrization. In *International Conference on Machine Learning*, 2021b.

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, Sergey Yekhanin, and Huishuai Zhang. Differentially private fine-tuning of language models. In *International Conference on Learning Representations*, 2022.

Chulhee Yun, Suvrit Sra, and Ali Jadbabaie. Small nonlinearities in activation functions create bad local minima in neural networks. In *International Conference on Learning Representations*, 2019a.

Chulhee Yun, Suvrit Sra, and Ali Jadbabaie. Small ReLU networks are powerful memorizers: a tight analysis of memorization capacity. In *Advances in Neural Information Processing Systems*, 2019b.

Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar. Are transformers universal approximators of sequence-to-sequence functions? In *International Conference on Learning Representations*, 2020.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires re-thinking generalization. In *International Conference on Learning Representations*, 2017a.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017b.

Chiyuan Zhang, Daphne Ippolito, Katherine Lee, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. Counterfactual memorization in neural language models. In *Advances in Neural Information Processing Systems*, 2023.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, 2019.

Jingzhao Zhang, Aditya Krishna Menon, Andreas Veit, Srinadh Bhojanapalli, Sanjiv Kumar, and Suvrit Sra. Coping with label shift via distributionally robust optimisation. In *International Conference on Learning Representations*, 2021.

Chunting Zhou, Xuezhe Ma, Paul Michel, and Graham Neubig. Examining and combating spurious features under distribution shift. In *International Conference on Machine Learning*, 2021a.

Yingxue Zhou, Steven Wu, and Arindam Banerjee. Bypassing the ambient dimension: Private sgd with gradient subspace identification. In *International Conference on Learning Representations*, 2021b.

Zhenyu Zhu, Fanghui Liu, Grigorios Chrysos, and Volkan Cevher. Robustness in deep learning: The good (width), the bad (depth), and the ugly (initialization). In *Advances in Neural Information Processing Systems*, 2022.

Indre Zliobaite. On the relation between accuracy and fairness in binary classification. In *2nd ICML Workshop on Fairness, Accountability, and Transparency in Machine Learning (FATML)*, 2015.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

Difan Zou and Quanquan Gu. An improved analysis of training over-parameterized deep neural networks. In *Advances in Neural Information Processing Systems*, 2019.

Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Gradient descent optimizes over-parameterized deep relu networks. *Machine Learning*, 109(3):467–492, 2020.

Guangyi Zou and Roman Vershynin. On the subgaussianity of quantized linear maps: An ai-assisted note. *arXiv preprint arXiv:2605.27563*, 2026.

APPENDIX A

Declaration of the use of generative AI

In this thesis, large language models such as ChatGPT, Gemini and Claude assisted in refining the writing style. Traditional data visualization methods were complemented with these same systems to enhance efficiency. The author acknowledges full responsibility for all included content.

APPENDIX B

Additional notation and remarks

Given a sub-exponential random variable X, let $\|X\|_{\psi_1} = \inf\{t > 0 \ : \ \mathbb{E}[\exp(|X|/t)] \leq 2\}$. Similarly, for a sub-Gaussian random variable, let $\|X\|_{\psi_2} = \inf\{t > 0 \ : \ \mathbb{E}[\exp(X^2/t^2)] \leq 2\}$. We use the analogous definitions for vectors. In particular, let $X \in \mathbb{R}^n$ be a random vector, then $\|X\|_{\psi_2} := \sup_{\|u\|_2=1} \left\|u^\top X\right\|_{\psi_2}$ and $\|X\|_{\psi_1} := \sup_{\|u\|_2=1} \left\|u^\top X\right\|_{\psi_1}$. Notice that if a vector has independent, mean-0, sub-Gaussian (sub-exponential) entries, then it is sub-Gaussian (sub-exponential). This is a direct consequence of Hoeffding's inequality and Bernstein's inequality (see Theorems 2.6.3 and 2.8.2 in Vershynin [2018]).

We use the definition of the Orlicz norm of order α of a real random variable X as

$$\|X\|_{\psi_\alpha} := \inf\{t > 0 \ : \ \mathbb{E}\left[\exp(|X|^\alpha/t^\alpha)\right] \leq 2\}. \tag{B.1}$$

From this definition, it follows that $\||X|^\gamma\|_{\psi_{\alpha/\gamma}} = \|X\|_{\psi_\alpha}^\gamma$.

We say that a random variable or vector respects the Lipschitz concentration property if there exists an absolute constant $c > 0$ such that, for every Lipschitz continuous function $\tau : \mathbb{R}^d \to \mathbb{R}$, we have $\mathbb{E}|\tau(X)| < +\infty$, and for all $t > 0$,

$$\mathbb{P}\left(|\tau(x) - \mathbb{E}_X[\tau(x)]| > t\right) \leq 2e^{-ct^2/\|\tau\|_{\text{Lip}}^2}. \tag{B.2}$$

When we state that a random variable or vector X is sub-Gaussian (or sub-exponential), we implicitly mean $\|X\|_{\psi_2} = O(1)$, *i.e.* it does not increase with the scalings of the problem. Notice that, if X is Lipschitz concentrated, then $X - \mathbb{E}[X]$ is sub-Gaussian. If $X \in \mathbb{R}$ is sub-Gaussian and $\tau : \mathbb{R} \to \mathbb{R}$ is Lipschitz, we have that $\tau(X)$ is sub-Gaussian as well. Also, if a random variable is sub-Gaussian or sub-exponential, its p-th momentum is upper bounded by a constant (that might depend on p).

We recall that the maximum of n Gaussian (not necessarily independent) random variables is smaller than $\log N$ with probability at least $1 - \exp(c \log^2 n)$, see, e.g., Section 1.4 of Rigollet and Hütter [2023].

In general, we indicate with C and c absolute, strictly positive, numerical constants, that do not depend on the scalings of the problem, *i.e.* input dimension, number of neurons, or number of training samples. Their value may change from line to line.

Given a matrix A, we indicate with $A_{i:}$ its i-th row, and with $A_{:j}$ its j-th column. Given a square matrix A, we denote by $\lambda_{\min}(A)$ its smallest eigenvalue. Given a matrix A, we indicate

with $\sigma_{\min}(A) = \sqrt{\lambda_{\min}(A^\top A)}$ its smallest singular value, with $\|A\|_{\text{op}}$ its operator norm (and largest singular value), and with $\|A\|_F$ its Frobenius norm ($\|A\|_F^2 = \sum_{ij} A_{ij}^2$).

Given two matrices $A, B \in \mathbb{R}^{m\times n}$, we denote by $A \circ B$ their Hadamard product, and by $A * B = [(A_{1:} \otimes B_{1:}), \dots, (A_{m:} \otimes B_{m:})]^\top \in \mathbb{R}^{m\times n^2}$ their row-wise Kronecker product (also known as Khatri-Rao product). We denote $A^{*2} = A * A$. We remark that $(A * B)(A * B)^\top = AA^\top \circ BB^\top$. We say that a matrix $A \in \mathbb{R}^{n\times n}$ is positive semi definite (p.s.d.) if it's symmetric and for every vector $v \in \mathbb{R}^n$ we have $v^\top Av \geq 0$.

Hermite polynomials

In this subsection, we refresh standard notions on the Hermite polynomials. For a more comprehensive discussion, we refer to O'Donnell [2014]. The (probabilist's) Hermite polynomials $\{h_j\}_{j\in\mathbb{N}}$ are an orthonormal basis for $L^2(\mathbb{R}, \gamma)$, where γ denotes the standard Gaussian measure. The following result holds.

Proposition B.1 (Proposition 11.31, O'Donnell [2014])**.** *Let ρ_1, ρ_2 be two standard Gaussian random variables, with correlation $\rho \in [-1, 1]$. Then,*

$$\mathbb{E}_{\rho_1,\rho_2}[h_i(\rho_1)h_j(\rho_2)] = \delta_{ij}\rho^i, \tag{B.3}$$

where $\delta_{ij} = 1$ if $i = j$, and 0 otherwise.

The first 5 Hermite polynomials are

$$h_0(\rho) = 1, \quad h_1(\rho) = \rho, \quad h_2(\rho) = \frac{\rho^2 - 1}{\sqrt{2}}, \quad h_3(\rho) = \frac{\rho^3 - 3\rho}{\sqrt{6}}, \quad h_4(\rho) = \frac{\rho^4 - 6\rho^2 + 3}{\sqrt{24}}. \tag{B.4}$$

Proposition B.2 (Definition 11.34, O'Donnell [2014])**.** *Every function $\phi \in L^2(\mathbb{R}, \gamma)$ is uniquely expressible as*

$$\phi(\rho) = \sum_{i\in\mathbb{N}} \mu_i^\phi h_i(\rho), \tag{B.5}$$

where the real numbers μ_i^ϕ's are called the Hermite coefficients of ϕ, and the convergence is in $L^2(\mathbb{R}, \gamma)$. More specifically,

$$\lim_{n\to+\infty} \left\| \left(\sum_{i=0}^{n} \mu_i^\phi h_i(\rho) \right) - \phi(\rho) \right\|_{L^2(\mathbb{R},\gamma)} = 0. \tag{B.6}$$

This readily implies the following result.

Proposition B.3. *Let ρ_1, ρ_2 be two standard Gaussian random variables with correlation $\rho \in [-1, 1]$, and let $\phi, \tau \in L^2(\mathbb{R}, \gamma)$. Then,*

$$\mathbb{E}_{\rho_1,\rho_2}[\phi(\rho_1)\tau(\rho_2)] = \sum_{i\in\mathbb{N}} \mu_i^\phi \mu_i^\tau \rho^i. \tag{B.7}$$

APPENDIX C

Appendix for Chapter 2

C.1 Some Useful Estimates

Lemma C.1. *Under Assumption 2.5, we have that*

$$n \cdot \log^8 N_{L-1} = o(N_{L-1}N_{L-2}), \tag{C.1}$$

$$N_{L-2} \cdot \log^2 n \cdot \log^2 N_{L-1} = o(N_{L-1}N_{L-2}), \tag{C.2}$$

$$N_{L-1} \cdot \log^2 n \cdot \log^2 N_{L-1} = o(N_{L-1}N_{L-2}), \tag{C.3}$$

$$n \cdot \log^2 N_{L-1} \cdot \log^4 n = o(N_{L-1}N_{L-2}). \tag{C.4}$$

Proof. We start by proving (C.1). If $N_{L-1} = O\left(n^2\right)$, then

$$n \cdot \log^8 N_{L-1} = O\left(n \cdot \log^8 n\right) = o(N_{L-1}N_{L-2}), \tag{C.5}$$

where the last passage follows from (2.4). Conversely, if $N_{L-1} = \Omega(n^2)$, then

$$n \cdot \log^8 N_{L-1} = O\left(\sqrt{N_{L-1}} \cdot \log^8 N_{L-1}\right) = o(N_{L-1}) = o(N_{L-1}N_{L-2}), \tag{C.6}$$

which concludes the proof of (C.1).

To obtain (C.2), we can exploit the second requirement of Assumption 2.5, which implies that $\log n = O\left(\log N_{L-1}\right)$. This readily implies (C.2). Notice that (C.3) naturally follows since $N_{L-1} = O\left(N_{L-2}\right)$ by Assumption 2.4.

Finally, to obtain (C.4), we write

$$n \cdot \log^2 N_{L-1} \cdot \log^4 n = O\left(n \cdot \log^6 N_{L-1}\right) = o(N_{L-1}N_{L-2}), \tag{C.7}$$

where in the first passage we use that $\log n = O\left(\log N_{L-1}\right)$ (from the second requirement of Assumption 2.5) and the last passage follows from (C.1). □

Lemma C.2 (Lipschitz constant of function of the features)**.** *For all $l \in [L-1]$, and for every Lipschitz function φ, we have*

$$\left\|\varphi(g_l(x))\right\|_{\mathrm{Lip}} = O\left(1\right), \tag{C.8}$$

with probability at least

$$1 - 2l \exp(-N_{L-1}), \tag{C.9}$$

over $(W_k)_{k=1}^l$. *We recall that* φ *is applied component-wise to* $g_l(x)$, *and* $\varphi(g_l(x)) : \mathbb{R}^d \to \mathbb{R}^{N_l}$ *is intended as a function of* x.

Proof. Note that $\varphi(g_l)$ is a composition of Lipschitz functions. Thus,

$$\begin{aligned} \|\varphi(g_l)\|_{\text{Lip}} &\leq \|\varphi\|_{\text{Lip}} \|g_l\|_{\text{Lip}} \leq \|\varphi\|_{\text{Lip}} \|W_l\|_{\text{op}} \prod_{k=1}^{l-1} \left(\|W_k\|_{\text{op}} \|\phi\|_{\text{Lip}} \right) \\ &\leq \|\varphi\|_{\text{Lip}} \|W_l\|_{\text{op}} M^{l-1} \prod_{k=1}^{l-1} \|W_k\|_{\text{op}}, \end{aligned} \tag{C.10}$$

where the last step is justified by Assumption 2.3.

Recall that, by the assumption on the initialization of the weights, $(W_k)_{i,j} \sim_{\text{i.i.d.}} \mathcal{N}(0, \beta_k^2/N_{k-1})$, for some constant β_k which does not depend on the layer widths. Then, by Theorem 4.4.5 of Vershynin [2018], we have that, for any $k \in [l]$,

$$\|W_k\|_{\text{op}} \leq C \frac{\beta_k}{\sqrt{N_{k-1}}} (\sqrt{N_{k-1}} + 2\sqrt{N_k}), \tag{C.11}$$

with probability at least $1 - 2\exp(-N_k)$, C being a numerical constant. By Assumption 2.4 on the topology of the network, we can rewrite this result as

$$\|W_k\|_{\text{op}} = O(1), \tag{C.12}$$

with probability at least $1 - 2\exp(-N_{L-1})$. To conclude, using a union bound over the layers up to layer l, we have that

$$\|\varphi(g_l)\|_{\text{Lip}} = O(1), \tag{C.13}$$

with probability at least $1 - 2l\exp(-N_{L-1})$ over $(W_k)_{k=1}^l$. □

Lemma C.3. *We have that*

$$\|D_L\|_{\text{op}} \leq \log N_{L-1}, \tag{C.14}$$

with probability at least $1 - 2\exp(-c\log^2 N_{L-1})$ *over* W_L, *where* c *is a numerical constant.*

Proof. Recall that $D_L = \text{diag}(W_L)$ contains on the diagonal N_{L-1} independent Gaussian random variables $(D_L)_{ii} \sim \mathcal{N}(0, \beta_L^2)$. Thus, for any $i \in [N_{L-1}]$,

$$\mathbb{P}(|(D_L)_{ii}| > \log N_{L-1}) < 2\exp(-\log^2 N_{L-1}/(2\beta_L^2)), \tag{C.15}$$

which gives

$$\begin{aligned} \mathbb{P}(\|D_L\|_{\text{op}} > \log N_{L-1}) =& \mathbb{P}(\max_{i\in[N_{L-1}]} |(D_L)_{ii}| > \log N_{L-1}) \\ \leq& N_{L-1}\mathbb{P}(|(D_L)_{11}| > \log N_{L-1}) \\ <& 2\exp(\log N_{L-1} - \log^2 N_{L-1}/(2\beta_L^2)) \\ <& 2\exp(-c\log^2 N_{L-1}), \end{aligned} \tag{C.16}$$

where the second step is a union bound on the entries of D_L. This gives the desired result. □

Lemma C.4. *We have that*

$$\|D_L\phi'(g_{L-1}(x))\|_{\text{Lip}} = \mathcal{O}(\log N_{L-1}), \tag{C.17}$$

with probability at least $1 - 2\exp(-c\log^2 N_{L-1}) - C\exp(-N_{L-1})$ *over* $(W_k)_{k=1}^L$. *We recall that* ϕ' *is applied component-wise to* $g_l(x)$, $D_L\phi'(g_{L-1}(x)) : \mathbb{R}^d \to \mathbb{R}^{N_{L-1}}$ *is intended as a function of* x, *and* c *is a numerical constant.*

Proof. We know by composition of Lipschitz functions that

$$\|D_L\phi'(g_{L-1})\|_{\text{Lip}} \leq \|D_L\|_{\text{op}} \|\phi'(g_{L-1})\|_{\text{Lip}}. \tag{C.18}$$

By Assumption 2.3, ϕ' is Lipschitz. Hence, by combining Lemma C.2 (where we use $\varphi = \phi'$) and Lemma C.3, the result follows. □

Lemma C.5 (Exponential tails of quadratic forms)**.** *Let* $x \sim P_X$. *Let* $u : \mathbb{R}^d \to \mathbb{R}^{d_u}$ *and* $v : \mathbb{R}^d \to \mathbb{R}^{d_v}$ *be mean-0 Lipschitz functions with respect to* x, *i.e.,* $\mathbb{E}_x[u(x)] = 0$, $\mathbb{E}_x[v(x)] = 0$, $\|u\|_{\text{Lip}} = c_1$ *and* $\|v\|_{\text{Lip}} = c_2$. *Let* U *be a* $d_u \times d_v$ *matrix, and*

$$\Gamma(x) = u(x)^\top U v(x) - \mathbb{E}_x\left[u(x)^\top U v(x)\right]. \tag{C.19}$$

Then,

$$\|\Gamma\|_{\psi_1} < CK^2 \|U\|_F. \tag{C.20}$$

where $K = \sqrt{c_1^2 + c_2^2}$, *and* C *is a numerical constant.*

Proof. Consider the function $z(x) : \mathbb{R}^d \to \mathbb{R}^{d_u+d_v}$ obtained by concatenating the vectors u and v, i.e.,

$$z(x) := [u(x), v(x)]^\top. \tag{C.21}$$

One can readily verify that $\|z\|_{\text{Lip}} \leq \sqrt{c_1^2 + c_2^2} := K$. Let us set

$$M = \frac{1}{2}\left(\begin{array}{c|c} 0 & U \\ \hline U^\top & 0 \end{array}\right). \tag{C.22}$$

Then, we have that

$$\Gamma = z^\top M z - \mathbb{E}_x\left[z^\top M z\right]. \tag{C.23}$$

Since x satisfies Assumption 2.2 and $z(x)$ is Lipschitz, in order to obtain tail bounds on Γ, we can apply the version of the Hanson-Wright inequality given by Theorem 2.3 in Adamczak [2015]:

$$\begin{aligned}
\mathbb{P}(|\Gamma| > t) =& \mathbb{P}(|z^\top M z - \mathbb{E}_x\left[z^\top M z\right]| > t) \\
<& 2\exp\left(-\frac{1}{C_1}\min\left(\frac{t^2}{K^4\|M\|_F^2}, \frac{t}{K^2\|M\|_{\text{op}}}\right)\right) \\
\leq& 2\exp\left(-\frac{1}{C_1}\min\left(\frac{t^2}{K^4\|U\|_F^2}, \frac{t}{K^2\|U\|_{\text{op}}}\right)\right),
\end{aligned} \tag{C.24}$$

where C_1 is a numerical constant, and in the last step we use that $\|M\|_{\text{op}} = \|U\|_{\text{op}}$ and $\|M\|_F^2 = \|U\|_F^2/2 \leq \|U\|_F^2$. Thus, by Lemma 5.5 of Soltanolkotabi et al. [2018], we conclude that

$$\|\Gamma\|_{\psi_1} < C_2 K^2 \|U\|_F, \tag{C.25}$$

for some numerical constant C_2, which gives the desired result. □

Lemma C.6. *Let $u \in \mathbb{R}^{d_u}$ and $v \in \mathbb{R}^{d_v}$ be two mean-0 sub-Gaussian vectors such that $\|u\|_{\psi_2} = c_1$ and $\|v\|_{\psi_2} = c_2$. Set $A_{uv} = \mathbb{E}\left[uv^\top\right]$. Then,*

$$\|A_{uv}\|_{\mathsf{op}} \leq C(c_1 + c_2)^2, \tag{C.26}$$

where C is a numerical constant.

Proof. Consider the vector

$$z := [u, v]^\top. \tag{C.27}$$

Then, z is sub-Gaussian and, by triangle inequality on the vectors $[u, 0]$ and $[0, v]$, we have that $\|z\|_{\psi_2} \leq c_1 + c_2$. Since u and v are mean-0, then z is also mean-0 and its covariance matrix can be written as $A_z := \mathbb{E}\left[zz^\top\right]$. Furthermore, we can show that

$$\|A_z\|_{\mathsf{op}} \leq C(c_1 + c_2)^2. \tag{C.28}$$

In fact, let w be the unitary eigenvector associated to the maximum eigenvalue of A_z. Then,

$$\|A_z\|_{\mathsf{op}} = w^\top A_z w = \mathbb{E}\left[w^\top z z^\top w\right] = \mathbb{E}\left[(w^\top z)^2\right]. \tag{C.29}$$

Furthermore, we have that

$$\|z\|_{\psi_2} := \sup_{w' \text{s.t.} \|w'\|_2 = 1} \left\|(w')^\top z\right\|_{\psi_2} \geq \left\|w^\top z\right\|_{\psi_2} \geq \frac{1}{C}\sqrt{\mathbb{E}\left[(w^\top z)^2\right]}, \tag{C.30}$$

where C is a numerical constant, and the last inequality comes from Eq. (2.15) of Vershynin [2018]. By combining (C.29) and (C.30) with $\|z\|_{\psi_2} \leq c_1 + c_2$, (C.28) readily follows.

Finally, we have that

$$A_z = \left(\begin{array}{c|c} A_u & A_{uv} \\ \hline A_{uv}^\top & A_v \end{array}\right), \tag{C.31}$$

where $A_u := \mathbb{E}\left[uu^\top\right]$ and $A_v := \mathbb{E}\left[vv^\top\right]$. As A_u and A_v are PSD, we have that

$$\|A_{uv}\|_{\mathsf{op}} \leq \|A_z\|_{\mathsf{op}}. \tag{C.32}$$

Hence, the desired result follows from (C.28) and (C.32). □

Lemma C.7. *Let A be an $n \times m$ matrix whose rows A_i are i.i.d. mean-0 sub-Gaussian random vectors in $\mathbb{R}^m$. Let $K = \|A_i\|_{\psi_2}$ the sub-Gaussian norm of each row. Then, we have*

$$\left\|AA^\top\right\|_{\mathsf{op}} = K^2 O\left(n + m\right), \tag{C.33}$$

with probability at least $1 - 2\exp(-cn)$, for some numerical constant c.

Proof. Without loss of generality, we can assume $K = 1$ to simplify the proof. Let Σ be the second moment matrix of each of the rows of A. Then, $\Sigma = \mathbb{E}\left[A_i A_i^\top\right]$, since the rows are mean-0. Note that, as the rows of A are i.i.d., Σ is independent of i. Furthermore, Lemma C.6 implies that the covariance matrix $\mathbb{E}\left[A_i A_i^\top\right]$ has operator norm bounded by a constant, since the sub-Gaussian norm of the rows is 1. Then, by using Remark 5.40 in Vershynin [2012], we have that

$$\left\|\frac{A^\top A}{n} - \Sigma\right\|_{\mathsf{op}} \leq \max(\delta, \delta^2), \qquad \text{where } \delta = C\sqrt{\frac{m}{n}} + \frac{t}{\sqrt{n}}, \tag{C.34}$$

with probability at least $1 - 2\exp(-ct^2)$, where c and C are numerical constants. Setting $t = \sqrt{m}$ and using a triangular inequality gives that, with probability at least $1 - 2\exp(-ct^2)$,

$$\left\|AA^\top\right\|_{\mathsf{op}} = \left\|A^\top A\right\|_{\mathsf{op}} \leq n \left\|\Sigma\right\|_{\mathsf{op}} + \max(C\sqrt{mn} + \sqrt{mn}, (C\sqrt{m} + \sqrt{m})^2) = O\left(n + m\right), \tag{C.35}$$

which implies the desired result (after re-scaling by K). □

Lemma C.8. *Let $\tilde{F}_l = F_l - \mathbb{E}_x[F_l] \in \mathbb{R}^{n \times N_l}$ be the centered features matrix at layer l. Then, we have*

$$\left\|\tilde{F}_l \tilde{F}_l^\top\right\|_{\mathsf{op}} = O\left(n + N_l\right), \tag{C.36}$$

with probability at least $1 - C\exp\left(-cN_{L-1}\right)$ over $(W_k)_{k=1}^l$ and $(x_i)_{i=1}^n \sim_{\text{i.i.d.}} P_X$.

Proof. From Lemma C.2, we have that

$$\left\|f_l(x)\right\|_{\text{Lip}} = \Theta(1), \tag{C.37}$$

with probability at least

$$1 - C'\exp\left(-N_{L-1}\right), \tag{C.38}$$

over $(W_k)_{k=1}^l$. We condition on this event in the rest of the proof.

Since $(x_i)_{i=1}^n \sim_{\text{i.i.d.}} P_X$ and P_X satisfies Assumption 2.2, all the rows of $\tilde{F}_l$ are mean-0 sub-Gaussian vectors, with sub-Gaussian norm bounded by a numerical constant. Here, we fix $(W_k)_{k=1}^l$ s.t. (C.37) holds, and the "mean-0" and the "sub-Gaussian norm" is intended w.r.t. the probability space of $(x_i)_{i=1}^n$.

An application of Lemma C.7 gives that

$$\left\|\tilde{F}_l \tilde{F}_l^\top\right\|_{\mathsf{op}} = O\left(n + N_l\right), \tag{C.39}$$

with probability at least $1 - 2\exp\left(-cN_l\right)$ over $(x_i)_{i=1}^n \sim_{\text{i.i.d.}} P_X$. Taking into account the previous conditioning, we conclude that

$$\left\|\tilde{F}_l \tilde{F}_l^\top\right\|_{\mathsf{op}} = O\left(n + N_l\right), \tag{C.40}$$

with probability at least $1 - (C' + 2)\exp\left(-cN_{L-1}\right)$ over $(W_k)_{k=1}^l$ and $(x_i)_{i=1}^n \sim_{\text{i.i.d.}} P_X$. □

Lemma C.9. *Let $\tilde{B}_{L-1} = B_{L-1} - \mathbb{E}_x[B_{L-1}] \in \mathbb{R}^{n \times N_{L-1}}$ be the centered back-propagation matrix at layer $L-1$. Then, we have*

$$\left\|\tilde{B}_{L-1} \tilde{B}_{L-1}^\top\right\|_{\mathsf{op}} = O\left(\left(n + N_{L-1}\right) \log^2 N_{L-1}\right), \tag{C.41}$$

with probability at least $1 - C\exp\left(-cN_{L-1}\right)$ over $(W_k)_{k=l}^L$ and $(x_i)_{i=1}^n \sim_{\text{i.i.d.}} P_X$.

Proof. From Lemma C.4, we have that

$$\left\|D_L \phi'(g_{L-1}(x))\right\|_{\text{Lip}} = \mathcal{O}(\log N_{L-1}), \tag{C.42}$$

with probability at least

$$1 - 2\exp(-c\log^2 N_{L-1}) - C\exp\left(-N_{L-1}\right), \tag{C.43}$$

over $(W_k)_{k=1}^L$. We condition on this event in the rest of the proof.

Since $(x_i)_{i=1}^n \sim_{\text{i.i.d.}} P_X$ and P_X satisfies Assumption 2.2, all the rows of $\tilde{B}_{L-1}$ are mean-0 sub-Gaussian vectors, with sub-Gaussian norm $\mathcal{O}(\log N_{L-1})$. Here, we fix $(W_k)_{k=1}^L$ s.t. (C.42) holds, and the "mean-0" and the "sub-Gaussian norm" is intended w.r.t. the probability space of $(x_i)_{i=1}^n$.

An application of Lemma C.7 gives that

$$\left\|\tilde{B}_{L-1}\tilde{B}_{L-1}^\top\right\|_{\mathsf{op}} = O\left((n+N_{L-1})\log^2 N_{L-1}\right), \tag{C.44}$$

with probability at least $1-2\exp\left(-cN_{L-1}\right)$ over $(x_i)_{i=1}^n \sim_{\text{i.i.d.}} P_X$. Taking into account the previous conditioning, we conclude that

$$\left\|\tilde{B}_{L-1}\tilde{B}_{L-1}^\top\right\|_{\mathsf{op}} = O\left((n+N_{L-1})\log^2 N_{L-1}\right), \tag{C.45}$$

with probability at least $1-(C'+2)\exp\left(-cN_{L-1}\right)$ over $(W_k)_{k=1}^L$ and $(x_i)_{i=1}^n \sim_{\text{i.i.d.}} P_X$. □

Lemma C.10. *Let ρ be a standard Gaussian random variable, then we have that*

$$\varphi_1(c) := \mathbb{E}_\rho\left[\phi(c\rho)\right], \tag{C.46}$$

and

$$\varphi_2(c) := \mathbb{E}_\rho\left[\phi^2(c\rho)\right], \tag{C.47}$$

are continuous functions in c. Furthermore, $\varphi_1(c)$ is Lipschitz in c, and

$$|\varphi_2(c_1)-\varphi_2(c_2)| \le C_1\,|c_1-c_2| + C_2|c_1^2-c_2^2|, \tag{C.48}$$

where C_1 and C_2 are numerical constants (independent of c_1, c_2).

Proof. Let $p(\rho)=\frac{1}{\sqrt{2\pi}}e^{-\rho^2/2}$. Then, we have

$$\begin{aligned}
|\varphi_1(c+\varepsilon)-\varphi_1(c)| &\le \int p(\rho)\,|\phi((c+\varepsilon)\rho)-\phi(c\rho)|\,d\rho \\
&\le \int p(\rho)\,|M\varepsilon\rho|\,d\rho \\
&= M\varepsilon\mathbb{E}_\rho\left[|\rho|\right] \\
&= C\varepsilon,
\end{aligned} \tag{C.49}$$

where in the second line we use that ϕ is M-Lipschitz by Assumption 2.3. Similarly, we have

$$\begin{aligned}
|\varphi_2(c+\varepsilon)-\varphi_2(c)| &\le \int p(\rho)\left|\phi^2((c+\varepsilon)\rho)-\phi^2(c\rho)\right|d\rho \\
&= \int p(\rho)\,|\phi((c+\varepsilon)\rho)-\phi(c\rho)|\,|\phi((c+\varepsilon)\rho)+\phi(c\rho)|\,d\rho \\
&\le \int p(\rho)\,|M\varepsilon\rho|\,(2\,|\phi(0)|+M(|c+\varepsilon|+|\varepsilon|)|\rho|)\,d\rho \\
&= C_1\varepsilon\mathbb{E}_\rho\left[|\rho|\right] + C_2\varepsilon|c|\mathbb{E}_\rho\left[\rho^2\right] + C_3\varepsilon^2\mathbb{E}_\rho\left[\rho^2\right] \\
&= C_4\varepsilon + C_2|c|\varepsilon + C_3\varepsilon^2 \\
&\le C_5\varepsilon + C_6\left|(c+\varepsilon)^2-c^2\right|.
\end{aligned} \tag{C.50}$$

□

Lemma C.11. *Let ρ be a standard Gaussian distribution, and $c \neq 0$ be an absolute constant. Then, we have*

$$|\mathbb{E}_\rho [\phi(c\rho)]| = O(1). \tag{C.51}$$

It also holds

$$|\mathbb{E}_\rho [\phi'(c\rho)]| = O(1). \tag{C.52}$$

Proof. For the first statement, we exploit the fact that ϕ is Lipschitz:

$$|\mathbb{E}_\rho [\phi(c\rho)]| \leq \mathbb{E}_\rho [|\phi(0)| + M|c\rho|] = |\phi(0)| + M|c|\mathbb{E}_\rho [|\rho|] = C_1. \tag{C.53}$$

The statement on ϕ' is easily derived following the same proof and using that $\|\phi'\|_{\text{Lip}} \leq M'$. □

Lemma C.12. *Let ρ be a standard Gaussian distribution, and $c \neq 0$ be an absolute constant. Then, we have*

$$\mathbb{E}_\rho \left[\phi^2(c\rho)\right] = \Theta(1), \tag{C.54}$$

and

$$\mathbb{E}_\rho \left[(\phi'(c\rho))^2\right] = \Theta(1). \tag{C.55}$$

Proof. For the upper-bound of the first statement, we exploit the fact that ϕ is Lipschitz:

$$\mathbb{E}_\rho \left[\phi^2(c\rho)\right] \leq \mathbb{E}_\rho \left[(|\phi(0)| + M|c\rho|)^2\right] = \phi^2(0) + 2M|c||\phi(0)|\mathbb{E}_\rho [|\rho|] + M^2c^2\mathbb{E}_\rho \left[\rho^2\right] = C_1. \tag{C.56}$$

For the lower bound, since ϕ is non-zero and continuous, we have that there exist a strictly positive constant $c' > 0$ and an interval $[c_1, c_2]$ with $c_2 > c_1$ such that $\phi^2(x) \geq c'$ for each $x \in [c_1, c_2]$. Therefore, we have

$$\mathbb{E}_\rho \left[\phi^2(c\rho)\right] \geq c'\mathbb{P}(c_1 \leq c\rho \leq c_2) = C_2. \tag{C.57}$$

The second statements is proved in the same way, as ϕ' is a non-zero Lipschitz function. □

Lemma C.13. *Let $\varphi : \mathbb{R}^d \to \mathbb{R}$ a Lipschitz function, and let $x \sim P_X$. Then,*

$$\mathbb{E}_x^2 [\varphi(x)] \geq \mathbb{E}_x \left[\varphi(x)^2\right] - c\|\varphi\|_{\text{Lip}}^2, \tag{C.58}$$

where c is a numerical constant.

Proof. We have

$$\begin{aligned}
\mathbb{E}_x^2 [\varphi(x)] &= \mathbb{E}_x \left[(\varphi(x))^2\right] - \mathbb{E}_x \left[(\varphi(x) - \mathbb{E}_x [\varphi(x)])^2\right] \\
&= \mathbb{E}_x \left[\varphi(x)^2\right] - \int_0^{+\infty} \mathbb{P}\left((\varphi(x) - \mathbb{E}_x [\varphi(x)])^2 > t\right) dt \\
&= \mathbb{E}_x \left[\varphi(x)^2\right] - \int_0^{+\infty} \mathbb{P}\left(|\varphi(x) - \mathbb{E}_x [\varphi(x)]| > \sqrt{t}\right) dt \\
&\geq \mathbb{E}_x \left[\varphi(x)^2\right] - \int_0^{+\infty} 2\exp\left(-Ct/\|\varphi\|_{\text{Lip}}^2\right) dt \\
&= \mathbb{E}_x \left[\varphi(x)^2\right] - 2\|\varphi\|_{\text{Lip}}^2 / C,
\end{aligned} \tag{C.59}$$

where the inequality is a consequence of Assumption 2.2. □

Lemma C.14. *Let $x \sim P_X$, and define $\tilde{c}_l(x) = \beta_l \|f_l(x)\| / \sqrt{N_l}$.*

Then, we have

$$\mathbb{E}_x \left[|\tilde{c}_l(x) - \mathbb{E}_x \left[\tilde{c}_l(x) \right]| \right] \leq C \frac{\|f_l\|_{\text{Lip}}}{\sqrt{N_l}}, \tag{C.60}$$

and

$$\mathbb{E}_x \left[\left(\tilde{c}_l(x) - \mathbb{E}_x \left[\tilde{c}_l(x) \right] \right)^2 \right] \leq C \frac{\|f_l\|_{\text{Lip}}^2}{N_l}, \tag{C.61}$$

where C is a numerical constant.

Proof. We have that

$$\begin{aligned}
\mathbb{E}_x \left[|\tilde{c}_l(x) - \mathbb{E}_x \left[\tilde{c}_l(x) \right]| \right] &= \int_0^{+\infty} \mathbb{P} \left(|\tilde{c}_l(x) - \mathbb{E}_x \left[\tilde{c}_l(x) \right]| > t \right) dt \\
&= \int_0^{+\infty} \mathbb{P} \left(|\|f_l(x)\| - \mathbb{E}_x \left[\|f_l(x)\| \right]| > \sqrt{N_l} t / \beta_l \right) dt \\
&\leq \int_0^{+\infty} 2 \exp \left(-c N_l t^2 / \|f_l\|_{\text{Lip}}^2 \right) dt \\
&= C \frac{\|f_l\|_{\text{Lip}}}{\sqrt{N_l}},
\end{aligned} \tag{C.62}$$

where c and C are numerical constants, and the third line is justified by Assumption 2.2.

Similarly, we have

$$\begin{aligned}
\mathbb{E}_x \left[\left(\tilde{c}_l(x) - \mathbb{E}_x \left[\tilde{c}_l(x) \right] \right)^2 \right] &= \int_0^{+\infty} \mathbb{P} \left(\left(\tilde{c}_l(x) - \mathbb{E}_x \left[\tilde{c}_l(x) \right] \right)^2 > t \right) dt \\
&= \int_0^{+\infty} \mathbb{P} \left(|\|f_l(x)\| - \mathbb{E}_x \left[\|f_l(x)\| \right]| > \sqrt{N_l t} / \beta_l \right) dt \\
&\leq \int_0^{+\infty} 2 \exp \left(-c N_l t / \|f_l\|_{\text{Lip}}^2 \right) dt \\
&= C \frac{\|f_l\|_{\text{Lip}}^2}{N_l},
\end{aligned} \tag{C.63}$$

where, again, c and C are numerical constants and the third line is justified by Assumption 2.2. □

Lemma C.15. *Let ρ_1 and ρ_2 be two standard Gaussian random variables, possibly correlated. Then, we have*

$$\begin{aligned}
&|\mathbb{E}_{\rho_1 \rho_2} \left[\phi(\rho_1 x_1) \phi(\rho_2 x_2) - \phi(\rho_1 y_1) \phi(\rho_2 y_2) \right]| \leq \\
&\qquad \leq C_1 |x_1 - y_1| + C_2 |x_2| |x_1 - y_1| + C_3 |x_2 - y_2| + C_4 |y_1| |x_2 - y_2| ,
\end{aligned} \tag{C.64}$$

where C_1, C_2, C_3, C_4 are numerical constants (which do not depend on x_1, x_2, y_1, y_2). Furthermore, the same result holds with ϕ' instead of ϕ.

Proof. We have

$$
\begin{aligned}
&\left|\mathbb{E}_{\rho_1\rho_2}\left[\phi(\rho_1 x_1)\phi(\rho_2 x_2) - \phi(\rho_1 y_1)\phi(\rho_2 y_2)\right]\right| \\
&\leq \left|\mathbb{E}_{\rho_1\rho_2}\left[\phi(\rho_1 x_1)\phi(\rho_2 x_2) - \phi(\rho_1 y_1)\phi(\rho_2 x_2)\right] + \mathbb{E}_{\rho_1\rho_2}\left[\phi(\rho_1 y_1)\phi(\rho_2 x_2) - \phi(\rho_1 y_1)\phi(\rho_2 y_2)\right]\right| \\
&\leq \mathbb{E}_{\rho_1\rho_2}\left[\left|\phi(\rho_1 x_1) - \phi(\rho_1 y_1)\right| \left|\phi(\rho_2 x_2)\right|\right] + \mathbb{E}_{\rho_1\rho_2}\left[\left|\phi(\rho_2 x_2) - \phi(\rho_2 y_2)\right| \left|\phi(\rho_1 y_1)\right|\right] \\
&\leq \mathbb{E}_{\rho_1\rho_2}\left[\left|M\rho_1(x_1 - y_1)\right| \left(\left|\phi(0)\right| + M\left|\rho_2 x_2\right|\right)\right] + \mathbb{E}_{\rho_1\rho_2}\left[\left|M\rho_2(x_2 - y_2)\right| \left(\left|\phi(0)\right| + M\left|\rho_1 y_1\right|\right)\right] \\
&\leq C_1\left|x_1 - y_1\right| \mathbb{E}\left[\left|\rho_1\right|\right] + C_2\left|x_2\right|\left|x_1 - y_1\right| \mathbb{E}\left[\left|\rho_1\right|\left|\rho_2\right|\right] \\
&\qquad + C_3\left|x_2 - y_2\right| \mathbb{E}\left[\left|\rho_2\right|\right] + C_4\left|y_1\right|\left|x_2 - y_2\right| \mathbb{E}\left[\left|\rho_1\right|\left|\rho_2\right|\right] \\
&\leq C_1\left|x_1 - y_1\right| + C_2\left|x_2\right|\left|x_1 - y_1\right| + C_3\left|x_2 - y_2\right| + C_4\left|y_1\right|\left|x_2 - y_2\right|,
\end{aligned}
\tag{C.65}
$$

where in third inequality we use that ϕ is M-Lipschitz, and in the last inequality we use that the quantities $\mathbb{E}\left[\left|\rho_1\right|\right]$, $\mathbb{E}\left[\left|\rho_2\right|\right]$ and $\mathbb{E}\left[\left|\rho_1\right|\left|\rho_2\right|\right]$ are all smaller than 1 (regardless of the correlation between ρ_1 and ρ_2). Since we only used the fact that ϕ is M-Lipschitz, the same result holds with ϕ' in place of ϕ.

□

C.2 Concentration of L2 Norms

In this appendix, we state and prove a number of high-probability estimates on the ℓ_2 norms of feature and backpropagation vectors. More specifically, our results can be summarized as follows:

- Lemma C.16 gives tight bounds on $\left\|f_l(x)\right\|_2$, *i.e.* the ℓ_2 norm of the feature vector at layer l. The statement holds with high probability over x and $(W_k)_{k=1}^l$.
- Lemmas C.17 and C.18 give tight bounds on $\mathbb{E}_x\left[\left\|f_l(x)\right\|_2^2\right]$ and $\mathbb{E}_x\left[\left\|f_l(x)\right\|_2\right]$, respectively. These quantities represent the *expectation* with respect to x of the (squared) ℓ_2 norm of the feature vector at layer l. The statements hold with high probability over $(W_k)_{k=1}^l$.
- Lemma C.19 focuses on the *centered* feature vector $f_l(x) - \mathbb{E}_x\left[f_l(x)\right]$, and it gives tight bounds on *(i)* its expected (w.r.t. x) squared ℓ_2 norm $\mathbb{E}_x\left[\left\|f_l(x) - \mathbb{E}_x\left[f_l(x)\right]\right\|_2^2\right]$, *(ii)* its expected (w.r.t. x) ℓ_2 norm $\mathbb{E}_x\left[\left\|f_l(x) - \mathbb{E}_x\left[f_l(x)\right]\right\|_2\right]$, and *(iii)* its ℓ_2 norm $\left\|f_l(x) - \mathbb{E}_x\left[f_l(x)\right]\right\|_2$. The first two statements hold with high probability over $(W_k)_{k=1}^l$, and the probability in the last statement is also over x.
- Lemma C.20 focuses on the *centered* backpropagation vector at layer $L-1$, and it gives tight bounds on its ℓ_2 norm $\left\|D_L\phi'(g_{L-1}(x)) - \mathbb{E}_x\left[D_L\phi'(g_{L-1}(x))\right]\right\|_2$. This statement holds with high probability over x and $(W_k)_{k=1}^l$.

Throughout this appendix, we always assume that P_X satisfies Assumptions 2.1 and 2.2, and that the layer widths satisfy Assumption 2.4. Furthermore, we use that the activation ϕ and its derivative ϕ' are Lipschitz (see Assumption 2.3).

Lemma C.16 (L2 norm of features)**.** *Let $x \sim P_X$. Then, for every $0 \leq l \leq L-1$,*

$$
\left\|f_l(x))\right\|_2 = \Theta(\sqrt{N_l}),
\tag{C.66}
$$

with probability at least $1 - C\exp(-cN_{L-1})$ *over* x *and* $(W_k)_{k=1}^l$. *As usual,* ϕ *is applied component-wise on* $g_l(x)$*, and* c *and* C *are numerical constants.*

Proof. We prove this by induction over l, and we start with the base case ($l = 0$). Recall that we have defined $f_0(x) := x$. As the ℓ_2 norm is a 1-Lipschitz function, by Assumption 2.2, we have that

$$\mathbb{P}\left(\left|\|x\|_2 - \mathbb{E}[\|x\|_2]\right| > t\right) \leq 2e^{-ct^2}. \tag{C.67}$$

Furthermore, Assumption 2.1 implies that $\mathbb{E}[\|x\|_2] = \Theta(\sqrt{d})$, hence setting $t = \mathbb{E}[\|x\|_2]/2$ in (C.67) proves the desired result for the base case (recalling that $N_{L-1} = O(d)$ by Assumption 2.4).

By inductive hypothesis, we have

$$\|f_{l-1}(x))\|_2 = \Theta(\sqrt{N_{l-1}}), \tag{C.68}$$

with probability at least $1 - C\exp(-cN_{L-1})$.

Define $\tilde{c} := \beta_l \|f_{l-1}(x)\|_2 / \sqrt{N_{l-1}}$. From now on, we condition on a realization of x and $(W_k)_{k=1}^{l-1}$ such that $\tilde{c} = \Theta(1)$. By (C.68), this happens with probability at least $1 - C\exp(-cN_{L-1})$.

To ease the notation, we use the shorthands $f := f_{l-1}(x)$ and $W := W_l$. Then, we can write

$$\|f_l(x)\|_2 = \left\|\phi(W^\top f)\right\|_2 = \sqrt{N_l}\sqrt{\frac{1}{N_l}\sum_{i=1}^{N_l}\phi^2((W^\top)_{i:}f)}. \tag{C.69}$$

Recall that $(W_l)_{i,j} \sim_{\text{i.i.d.}} \mathcal{N}(0, \beta_l^2/N_{l-1})$ and that the Gaussian distribution is rotationally invariant. Thus, the RHS of (C.69) has the same distribution as

$$\sqrt{N_l}\sqrt{\frac{1}{N_l}\sum_{i=1}^{N_l}\phi^2\left(\tilde{c}\rho_i\right)} = \sqrt{N_l}\sqrt{\mathbb{E}_{\rho_1}\left[\phi^2\left(\tilde{c}\rho_1\right)\right] + \frac{1}{N_l}\sum_{i=1}^{N_l} Z_i}, \tag{C.70}$$

where $(\rho_i)_{i=1}^{N_l} \sim_{\text{i.i.d.}} \mathcal{N}(0,1)$ and also independent of f, and we have defined the independent, mean-0 random variables

$$Z_i = \phi^2\left(\tilde{c}\rho_i\right) - \mathbb{E}_{\rho_1}\left[\phi^2\left(\tilde{c}\rho_1\right)\right]. \tag{C.71}$$

Note that, in the definition of Z_i, the randomness comes only from ρ_i, since we are conditioning on $\tilde{c}$.

We have that

$$\left\|\phi\left(\tilde{c}\rho_i\right)\right\|_{\psi_2} \leq \left\|\phi\left(\tilde{c}\rho_i\right) - \mathbb{E}_{\rho_i}\left[\phi\left(\tilde{c}\rho_i\right)\right]\right\|_{\psi_2} + \left\|\mathbb{E}_{\rho_i}\left[\phi\left(\tilde{c}\rho_i\right)\right]\right\|_{\psi_2} \leq C_1 + C_2 = C_3, \tag{C.72}$$

where the first term is bounded by a constant by Theorem 5.2.2 in Vershynin [2018], and the bound on the second term follows from Lemma C.11. As a consequence, we have

$$\begin{aligned}
\left\|Z_i\right\|_{\psi_1} =& \left\|\phi^2\left(\tilde{c}\rho_i\right) - \mathbb{E}_{\rho_i}\left[\phi^2\left(\tilde{c}\rho_i\right)\right]\right\|_{\psi_1} \\
\leq& C_4\left\|\phi^2\left(\tilde{c}\rho_i\right)\right\|_{\psi_1} \\
=& C_4\left\|\phi\left(\tilde{c}\rho_i\right)\right\|_{\psi_2}^2 \\
\leq& C_5,
\end{aligned} \tag{C.73}$$

where the inequality in the second line follows from Exercise 2.7.10 of Vershynin [2018], the equality in the third line follows from Lemma 2.7.6 of Vershynin [2018], and the inequality in the last line follows from (C.72). Hence, the Z_i-s are i.i.d. sub-exponential random variables, with sub-exponential norm bounded by a numerical constant. An application of Bernstein inequality (cf. Corollary 2.8.3. in Vershynin [2018]) gives that

$$\mathbb{P}\left(\left|\frac{1}{N_l}\sum_{i=1}^{N_l} Z_i\right| > t\right) \leq 2\exp\left(-c\min\left(\frac{t^2}{C_6^2}, \frac{t}{C_6}\right) N_l\right), \tag{C.74}$$

where c, C_6 are numerical constants. Furthermore, by Lemma C.12, we have

$$\mathbb{E}_{\rho_1}\left[\phi^2\left(\tilde{c}\rho_1\right)\right] = \Theta(1). \tag{C.75}$$

By setting $t = \mathbb{E}_{\rho_1}\left[\phi^2\left(\tilde{c}\rho_1\right)\right]/2$ into (C.74) and using (C.70) and (C.75), we conclude that

$$\left\|\phi(W^\top f)\right\|_2 = \Theta(\sqrt{N_l}), \tag{C.76}$$

with probability at least $1 - C\exp(-cN_{L-1}) - 2\exp(-cN_l) \geq 1 - C_1\exp(-cN_{L-1})$, for some numerical constant c and C_1, which concludes the proof. □

Lemma C.17 (Expected squared L2 norm of features)**.** *Let $x \sim P_X$. Then, for every $0 \leq l \leq L-1$,*

$$\mathbb{E}_x\left[\|f_l(x)\|_2^2\right] = \Theta(N_l), \tag{C.77}$$

with probability at least $1 - C\exp(-cN_{L-1})$ over $(W_k)_{k=1}^l$. As usual, c and C are numerical constants.

Proof. The argument is by induction over l. The base case is a direct consequence of Assumption 2.1, since $f_0(x) = x$.

By inductive hypothesis, we have

$$\mathbb{E}_x\left[\|f_{l-1}(x))\|_2^2\right] = \Theta(N_{l-1}), \tag{C.78}$$

with probability at least $1 - C\exp(-cN_{L-1})$. Define $\tilde{c}(x) := \beta_l \|f_{l-1}(x)\|_2 / \sqrt{N_{l-1}}$. From now on, we condition on a realization of $(W_k)_{k=1}^{l-1}$ such that $\mathbb{E}_x\left[\tilde{c}^2(x)\right] = \Theta(1)$. By (C.78), this happens with probability at least $1 - C\exp(-cN_{L-1})$.

To ease the notation, we use the shorthands $f := f_{l-1}(x)$, $W := W_l$ and $w_i = W_{:i}$. Then, we can write

$$\begin{aligned}
\mathbb{E}_x\left[\|f_l(x))\|_2^2\right] &= \mathbb{E}_x\left[\left\|\phi(W^\top f)\right\|_2^2\right] \\
&= N_l\left(\frac{1}{N_l}\sum_{i=1}^{N_l}\mathbb{E}_x\left[\phi^2((W^\top)_{i:}f)\right]\right) \\
&= N_l\left(\mathbb{E}_{w_1}\mathbb{E}_x\left[\phi^2(w_1^\top f)\right] + \frac{1}{N_l}\sum_{i=1}^{N_l} Z_i\right),
\end{aligned} \tag{C.79}$$

where we use that the w_i-s are equally distributed and we have defined the independent, mean-0 random variables

$$Z_i = \mathbb{E}_x\left[\phi^2\left(w_i^\top f(x)\right)\right] - \mathbb{E}_{w_i}\mathbb{E}_x\left[\phi^2\left(w_i^\top f(x)\right)\right]. \tag{C.80}$$

Note that, in the definition of Z_i, the randomness comes only from w_i, since we are conditioning on $(W_k)_{k=1}^{l-1}$.

We have that

$$\begin{aligned}\|Z_i\|_{\psi_1} \leq & \mathbb{E}_x\left[\left\|\phi^2\left(w_i^\top f(x)\right) - \mathbb{E}_{w_i}\left[\phi^2\left(w_i^\top f(x)\right)\right]\right\|_{\psi_1}\right] \\ \leq & \mathbb{E}_x\left[C_1\left\|\phi^2\left(w_i^\top f(x)\right)\right\|_{\psi_1}\right] \\ = & C_1\mathbb{E}_x\left[\left\|\phi\left(w_i^\top f(x)\right)\right\|_{\psi_2}^2\right],\end{aligned} \tag{C.81}$$

where the first line follows from Jensen's inequality as $\|\cdot\|_{\psi_1}$ is convex, the inequality in the second line follows from Exercise 2.7.10 of Vershynin [2018], and the equality in the third line follows from Lemma 2.7.6 of Vershynin [2018].

Recall that $(W_l)_{i,j} \sim_{\text{i.i.d.}} \mathcal{N}(0, \beta_l^2/N_{l-1})$ and that the Gaussian distribution is rotationally invariant. Thus, $\phi\left(w_i^\top f(x)\right)$ has the same distribution as $\phi\left(\tilde{c}(x)\rho_i\right)$, where $(\rho_i)_{i=1}^{N_l} \sim_{\text{i.i.d.}} \mathcal{N}(0,1)$ and also independent of $\tilde{c}(x)$. We now condition on a realization of x and $(W_k)_{k=1}^{l-1}$ and provide an upper bound on the sub-Gaussian norm $\left\|\phi\left(w_i^\top f(x)\right)\right\|_{\psi_2}$, where the only randomness comes again from w_i (and, hence, from ρ_i). We have that

$$\begin{aligned}\left\|\phi\left(w_i^\top f(x)\right)\right\|_{\psi_2} &= \left\|\phi\left(\tilde{c}(x)\rho_i\right)\right\|_{\psi_2} \\ &\leq \left\|\phi\left(\tilde{c}(x)\rho_i\right) - \mathbb{E}_{\rho_i}\left[\phi\left(\tilde{c}(x)\rho_i\right)\right]\right\|_{\psi_2} + \left\|\mathbb{E}_{\rho_i}\left[\phi\left(\tilde{c}(x)\rho_i\right)\right]\right\|_{\psi_2} \\ &\leq C_1\tilde{c}(x) + C_2\tilde{c}(x) + C_3 = C_4(\tilde{c}(x)+1).\end{aligned} \tag{C.82}$$

where the first term in the RHS in the second line is bounded by $C_1\tilde{c}(x)$ by Theorem 5.2.2 in Vershynin [2018], and the second term is bounded by $C_2\tilde{c}(x) + C_3$ by following the same proof of Lemma C.11 as ϕ is Lipschitz. By combining (C.81) and (C.82), we get

$$\|Z_i\|_{\psi_1} \leq C_4^2\mathbb{E}_x\left[(\tilde{c}(x)+1)^2\right] \leq C_5, \tag{C.83}$$

where we use that $\mathbb{E}_x\left[\tilde{c}^2(x)\right] = \Theta(1)$.

Hence, the Z_i-s are i.i.d. sub-exponential random variables, with sub-exponential norm bounded by a numerical constant. An application of Bernstein inequality (cf. Corollary 2.8.3. in Vershynin [2018]) gives that

$$\mathbb{P}\left(\left|\frac{1}{N_l}\sum_{i=1}^{N_l} Z_i\right| > t\right) \leq 2\exp\left(-c\min\left(\frac{t^2}{C_5^2}, \frac{t}{C_5}\right)N_l\right), \tag{C.84}$$

where c, C_5 are numerical constants.

Let us consider the first term in (C.79):

$$\mathbb{E}_x\mathbb{E}_{w_1}\left[\phi^2(w_1^\top f)\right] = \mathbb{E}_x\mathbb{E}_{\rho_1}\left[\phi^2\left(\tilde{c}(x)\rho_1\right)\right], \tag{C.85}$$

where the equality comes again from the rotational invariance of the Gaussian distribution of w_1. We will show that

$$\mathbb{E}_x\mathbb{E}_{\rho_1}\left[\phi^2\left(\tilde{c}(x)\rho_1\right)\right] = \Theta(1), \tag{C.86}$$

with probability at least $1 - C\exp(-cN_{L-1})$ over $(W_k)_{k=1}^{l-1}$.

The upper bound in (C.86) follows from the same passages in (C.56), as $\mathbb{E}_x[\tilde{c}^2(x)] = \Theta(1)$. We now prove the lower bound. By Lemma C.16, we have that there exist numerical constants $c_2 > c_1 > 0$ such that $\tilde{c}(x) \in [c_1, c_2]$ with probability at least $1 - C\exp(-cN_{L-1})$ over x and $(W_k)_{k=1}^{l-1}$. Hence, with probability at least $1 - 2C\exp(-cN_{L-1})$ over $(W_k)_{k=1}^{l-1}$, we have that

$$\mathbb{P}_x(\tilde{c}(x) \in [c_1, c_2]) \geq 1/2, \tag{C.87}$$

where we use the symbol $\mathbb{P}_x$ to highlight that this last probability is taken over x. Let us condition on a realization of $(W_k)_{k=1}^{l-1}$ s.t. (C.87) holds. Then, we have

$$\mathbb{E}_x\mathbb{E}_{\rho_1}\left[\phi^2\left(\tilde{c}(x)\rho_1\right)\right] \geq \frac{1}{2}\inf_{c\in[c_1,c_2]}\mathbb{E}_{\rho_1}\left[\phi^2\left(c\rho_1\right)\right]. \tag{C.88}$$

By Lemma C.10, we have that $\varphi(c) = \mathbb{E}_{\rho_1}\left[\phi^2\left(c\rho_1\right)\right]$ is continuous in c. Therefore, by Weierstrass theorem, there exists a strictly positive $c^* \in [c_1, c_2]$ such that $\inf_{c\in[c_1,c_2]}\mathbb{E}_{\rho_1}\left[\phi^2\left(c\rho_1\right)\right] = \mathbb{E}_{\rho_1}\left[\phi^2\left(c^*\rho_1\right)\right]$. Thus,

$$\mathbb{E}_x\mathbb{E}_{\rho_1}\left[\phi^2\left(\tilde{c}(x)\rho_1\right)\right] \geq \frac{1}{2}\mathbb{E}_{\rho_1}\left[\phi^2\left(c^*\rho_1\right)\right] = \Theta(1), \tag{C.89}$$

where the last equality is a consequence of Lemma C.12. This concludes the proof of the lower bound in (C.86).

By setting $t = \mathbb{E}_{\rho_1}\left[\phi^2\left(c^*\rho_1\right)\right]/4$ into (C.84) and using (C.86) and (C.79), we conclude that

$$\mathbb{E}_x\left[\|f_l(x))\|_2^2\right] = \Theta(N_l), \tag{C.90}$$

with probability at least $1 - C\exp(-cN_{L-1})$, for some numerical constants C and c, which concludes the proof. □

Lemma C.18 (Expected L2 norm of features)**.** *Let $x \sim P_X$. Then, for every $0 \leq l \leq L-1$,*

$$\mathbb{E}_x\left[\|f_l(x)\|_2\right] = \Theta(\sqrt{N_l}), \tag{C.91}$$

with probability at least $1 - C\exp(-cN_{L-1})$ over $(W_k)_{k=1}^{l}$. As usual, c and C are numerical constants.

Proof. We condition on the events

$$\mathbb{E}_x\left[\|f_l(x)\|_2^2\right] = \Theta(N_l), \tag{C.92}$$

and

$$\left\|\|f_l(x)\|_2\right\|_{\text{Lip}} = O\left(1\right), \tag{C.93}$$

which happen with probability at least $1 - C\exp(-cN_{L-1})$ over $(W_k)_{k=1}^{l}$ by Lemma C.17 and C.2.

The upper bound is a direct consequence of Jensen's inequality:

$$\mathbb{E}_x\left[\|f_l(x)\|_2\right] \leq \sqrt{\mathbb{E}_x\left[\|f_l(x)\|_2^2\right]} = \Theta(\sqrt{N_l}). \tag{C.94}$$

For the lower bound, we use Lemma C.13, and we obtain

$$\mathbb{E}_x\left[\|f_l(x)\|_2\right] \geq \sqrt{\mathbb{E}_x\left[\|f_l(x)\|_2^2\right] - c\left\|\|f_l(x)\|_2\right\|_{\text{Lip}}^2} = \Theta(\sqrt{N_l}). \tag{C.95}$$

□

Lemma C.19 (L2 norms of centered features)**.** *Let $x \sim P_X$. Then, for every $0 \leq l \leq L-1$, the following results hold.*

1. *With probability at least $1 - C\exp(-cN_{L-1})$ over $(W_k)_{k=1}^l$, we have that*

$$\mathbb{E}_x\left[\|f_l(x) - \mathbb{E}_x\left[f_l(x)\right]\|_2^2\right] = \Theta(N_l). \tag{C.96}$$

2. *With probability at least $1 - C\exp(-cN_{L-1})$ over $(W_k)_{k=1}^l$, we have that*

$$\mathbb{E}_x\left[\|f_l(x) - \mathbb{E}_x\left[f_l(x)\right]\|_2\right] = \Theta(\sqrt{N_l}). \tag{C.97}$$

3. *With probability at least $1 - C\exp(-cN_{L-1})$ over $(W_k)_{k=1}^l$ and x, we have that*

$$\|f_l(x) - \mathbb{E}_x\left[f_l(x)\right]\|_2 = \Theta(\sqrt{N_l}). \tag{C.98}$$

Proof. The argument is by induction over l. The base case for (C.96) follows directly from Assumption 2.1, since $f_0(x) = x$. Since the ℓ_2 norm is a 1-Lipschitz function, from Jensen inequality and Lemma C.13 we readily obtain the base case for (C.97). Note that $\|\|x - \mathbb{E}_x[x]\|_2\|_{\text{Lip}} \leq 1$. Then, the base case for (C.98) is a direct consequence of Assumption 2.2 on x and of the base case of (C.97), and it holds with probability at least $1 - C'\exp(-cd) \geq 1 - C'\exp(-cN_{L-1})$ over x.

By inductive hypothesis, we assume the three statements to be true for layer $l-1$, for $l \in [L-1]$. We will now prove (C.96) for layer l.

To ease the notation, we use the shorthands $f(x) := f_{l-1}(x)$, $f = \mathbb{E}_x[f(x)]$, $\tilde{f}(x) = f(x) - f$, $W := W_l$ and $w_i = W_{:i}$. We also define $\tilde{c}(x) = \beta_l \|f(x)\| / \sqrt{N_{l-1}}$ and $\tilde{c} = \mathbb{E}_x[\tilde{c}(x)]$.

We condition on the following events in the probability space of $(W_k)_{k=1}^{l-1}$:

(a) $\|f(x)\|_{\text{Lip}} = O(1)$, which happens with probability at least $1 - C'\exp(-cN_{L-1})$ by Lemma C.2.

(b) $\tilde{c} = \Theta(1)$, which happens with probability at least $1 - C'\exp(-cN_{L-1})$ by Lemma C.18. Notice that by Jensen inequality this also implies $\|f\|_2^2 = O(N_{l-1})$.

(c) By inductive hypothesis, we have that, with probability at least $1 - C'\exp(-cN_{n-1})$ over x and $(W_k)_{k=1}^{l-1}$,

$$\left\|\tilde{f}(x)\right\|_2^2 = \Theta(N_{l-1}). \tag{C.99}$$

Hence, with probability at least $1 - 2C'\exp(-cN_{n-1})$ over $(W_k)_{k=1}^{l-1}$, we have that

$$\mathbb{P}_x\left(c_1 N_{l-1} \leq \left\|\tilde{f}(x)\right\|_2^2 \leq c_2 N_{l-1}\right) \geq 1/2, \tag{C.100}$$

for some numerical constants $c_2 > c_1 > 0$. In (C.100), we use the symbol $\mathbb{P}_x$ to highlight that this last probability is taken over x. For the rest of the argument, we condition on a realization of $(W_k)_{k=1}^{l-1}$ s.t. (C.100) holds.

By taking a union bound, the events (a)-(c) happen with probability at least $1-4C'\exp(-cN_{n-1})$ over $(W_k)_{k=1}^{l-1}$.

Now, we can write

$$\begin{aligned}\mathbb{E}_x\left[\|f_l(x)-\mathbb{E}_x\left[f_l(x)\right]\|_2^2\right] &= \mathbb{E}_x\left[\left\|\phi(W^\top f(x))-\mathbb{E}_x\left[\phi(W^\top f(x))\right]\right\|_2^2\right]\\ &= N_l\left(\frac{1}{N_l}\sum_{i=1}^{N_l}\mathbb{E}_x\left[\left(\phi((W^\top)_{i:}f(x))-\mathbb{E}_x\left[\phi((W^\top)_{i:}f(x))\right]\right)^2\right]\right)\\ &= N_l\left(\mathbb{E}_{xw_1}\left[\left(\phi(w_1^\top f(x))-\mathbb{E}_x\left[\phi(w_1^\top f(x))\right]\right)^2\right]+\frac{1}{N_l}\sum_{i=1}^{N_l}Z_i\right),\end{aligned} \tag{C.101}$$

where we use that the w_i-s are identically distributed and we have defined the independent, mean-0 random variables

$$Z_i=\mathbb{E}_x\left[\left(\phi(w_i^\top f(x))-\mathbb{E}_x\left[\phi(w_i^\top f(x))\right]\right)^2\right]-\mathbb{E}_{w_1x}\left[\left(\phi(w_1^\top f(x))-\mathbb{E}_x\left[\phi(w_1^\top f(x))\right]\right)^2\right]. \tag{C.102}$$

As in the proof of Lemma C.17, in the definition of Z_i, the randomness comes only from w_i, since we are conditioning on $(W_k)_{k=1}^{l-1}$.

We have that

$$\begin{aligned}\|Z_i\|_{\psi_1} &\le C_0\left\|\mathbb{E}_x\left[\left(\phi(w_i^\top f(x))-\mathbb{E}_x\left[\phi(w_i^\top f(x))\right]\right)^2\right]\right\|_{\psi_1}\\ &\le C_0\mathbb{E}_x\left[\left\|\left(\phi(w_i^\top f(x))-\mathbb{E}_x\left[\phi(w_i^\top f(x))\right]\right)^2\right\|_{\psi_1}\right]\\ &= C_0\mathbb{E}_x\left[\left\|\phi(w_i^\top f(x))-\mathbb{E}_x\left[\phi(w_i^\top f(x))\right]\right\|_{\psi_2}^2\right]\\ &\le C_0\mathbb{E}_x\left[\left(\left\|\phi(w_i^\top f(x))\right\|_{\psi_2}+\left\|\mathbb{E}_x\left[\phi(w_i^\top f(x))\right]\right\|_{\psi_2}\right)^2\right]\\ &\le C_0\mathbb{E}_x\left[\left(\left\|\phi(w_i^\top f(x))\right\|_{\psi_2}+\mathbb{E}_x\left[\left\|\phi(w_i^\top f(x))\right\|_{\psi_2}\right]\right)^2\right],\end{aligned} \tag{C.103}$$

where C_0 is a numerical constant, the first inequality follows from Exercise 2.7.10 of Vershynin [2018], the second line follows from Jensen's inequality as $\|\cdot\|_{\psi_1}$ is convex, the equality follows from Lemma 2.7.6 of Vershynin [2018], and the last line follows from Jensen's inequality as $\|\cdot\|_{\psi_2}$ is convex.

Recall that $(W)_{i,j}\sim_{\text{i.i.d.}}\mathcal{N}(0,\beta_l^2/N_{L-1})$ and that the Gaussian distribution is rotationally invariant. Thus, $\phi'\left(w_i^\top f(x)\right)$ has the same distribution as $\phi'\left(\tilde{c}(x)\rho_i\right)$, where $(\rho_i)_{i=1}^{N_{L-1}}\sim_{\text{i.i.d.}}\mathcal{N}(0,1)$ and also independent of $\tilde{c}(x)$. Therefore,

$$\begin{aligned}\left\|\phi\left(w_i^\top f(x)\right)\right\|_{\psi_2} &= \left\|\phi\left(\tilde{c}(x)\rho_i\right)\right\|_{\psi_2}\\ &\le \left\|\phi\left(\tilde{c}(x)\rho_i\right)-\mathbb{E}_{\rho_i}\left[\phi\left(\tilde{c}(x)\rho_i\right)\right]\right\|_{\psi_2}+\left\|\mathbb{E}_{\rho_i}\left[\phi\left(\tilde{c}(x)\rho_i\right)\right]\right\|_{\psi_2}\\ &\le C_1\tilde{c}(x)+C_2\tilde{c}(x)+C_3\le C_4(\tilde{c}(x)+1),\end{aligned} \tag{C.104}$$

where the first term in the RHS in the second line is bounded by $C_1\tilde{c}(x)$ for Theorem 5.2.2 in Vershynin [2018], and the second term is bounded by $C_2\tilde{c}(x)+C_3$ by following the same proof of Lemma C.11.

Merging together (C.104) and (C.103) we get

$$\begin{aligned}\|Z_i\|_{\psi_1} \leq & C_0 \mathbb{E}_x \left[\left(C_4 \tilde{c}(x) + C_4 \tilde{c} + C_5 \right)^2 \right] \\ = & C_6 \mathbb{E}_x [(\tilde{c}(x) - \tilde{c})^2] + C_4 \tilde{c}^2 + C_7 \tilde{c} + C_8 \\ = & C_6 O\left(N_l^{-1} \right) + C_4 \tilde{c}^2 + C_7 \tilde{c} + C_8 \\ \leq & C_9,\end{aligned} \tag{C.105}$$

where in the third line we use Lemma C.14.

Hence, the Z_i-s are i.i.d. sub-exponential random variables, with sub-exponential norm bounded by a numerical constant. An application of Bernstein inequality (cf. Corollary 2.8.3. in Vershynin [2018]) gives that

$$\mathbb{P}\left(\left| \frac{1}{N_l} \sum_{i=1}^{N_l} Z_i \right| > t \right) \leq 2 \exp\left(-c \min\left(\frac{t^2}{C^2}, \frac{t}{C} \right) N_l \right), \tag{C.106}$$

where c, C are numerical constants. We recall that this probability is intended over W_l.

Let's now focus on the first term in the last line of (C.101), using the shorthand $w = w_1$, to ease the notation. We can rewrite this term as

$$\begin{aligned}&\mathbb{E}_w \left[\mathbb{E}_x \left[\phi^2(w^\top f(x)) \right] - \mathbb{E}_{xy} \left[\phi(w^\top f(x)) \phi(w^\top f(y)) \right] \right] \\ &\quad = \mathbb{E}_x \mathbb{E}_w \left[\phi^2(w^\top f(x)) \right] - \mathbb{E}_{xy} \mathbb{E}_w \left[\phi(w^\top f(x)) \phi(w^\top f(y)) \right].\end{aligned} \tag{C.107}$$

The aim of this part of the proof is to show that the quantity in (C.107) is $\Theta(1)$.

Recall that $(W_l)_{i,j} \sim_{\text{i.i.d.}} \mathcal{N}(0, \beta_l^2 / N_{l-1})$ and that the Gaussian distribution is rotationally invariant. Thus, $\phi\left(w^\top f(x) \right)$ has the same distribution as $\phi\left(\tilde{c}(x) \rho \right)$, where $\rho \sim \mathcal{N}(0,1)$ is independent of $\tilde{c}(x)$. We therefore have

$$\begin{aligned}&\left| \mathbb{E}_x \mathbb{E}_w \left[\phi^2\left(w^\top f(x) \right) - \phi^2\left(w^\top f(x) \frac{\tilde{c}}{\tilde{c}(x)} \right) \right] \right| \\ &= \left| \mathbb{E}_x \mathbb{E}_\rho \left[\phi^2(\rho \tilde{c}(x)) - \phi^2(\rho \tilde{c}) \right] \right| \\ &\leq \mathbb{E}_x \left[C_1 |\tilde{c} - \tilde{c}(x)| + C_2 \left| \tilde{c}^2 - \tilde{c}(x)^2 \right| \right] \\ &\leq \mathbb{E}_x \left[C_1 |\tilde{c} - \tilde{c}(x)| + C_2 (\tilde{c} - \tilde{c}(x))^2 + 2 C_2 \tilde{c} \, |\tilde{c} - \tilde{c}(x)| \right] \\ &= O\left(N_{l-1}^{-1/2} \right),\end{aligned} \tag{C.108}$$

where the third line follows from Lemma C.10, and the last passage follows from Lemma C.14. Similarly, we have

$$\begin{aligned}&\left| \mathbb{E}_{xy} \mathbb{E}_w \left[\phi(w^\top f(x)) \phi(w^\top f(y)) - \phi\left(w^\top f(x) \frac{\tilde{c}}{\tilde{c}(x)} \right) \phi\left(w^\top f(y) \frac{\tilde{c}}{\tilde{c}(y)} \right) \right] \right| \\ &= \left| \mathbb{E}_{xy} \mathbb{E}_{\rho_1 \rho_2} \left[\phi(\rho_1 \tilde{c}(x)) \phi(\rho_2 \tilde{c}(y)) - \phi(\rho_1 \tilde{c}) \phi(\rho_2 \tilde{c}) \right] \right| \\ &\leq \mathbb{E}_{xy} \left[C_1 \left| \tilde{c}(x) - \tilde{c} \right| + C_2 \tilde{c}(y) \left| \tilde{c}(x) - \tilde{c} \right| + C_3 \left| \tilde{c}(y) - \tilde{c} \right| + C_4 \tilde{c} \left| \tilde{c}(y) - \tilde{c} \right| \right] \\ &= O\left(N_{l-1}^{-1/2} \right),\end{aligned} \tag{C.109}$$

where ρ_1 and ρ_2 indicate two standard Gaussian random variables with correlation $f(x)^\top f(y) / (\|f(x)\| \, \|f(y)\|)$, the third line follows from C.15, and the last passage follows from Lemma C.14. By combining (C.108) and (C.109), we have that

$$\left| \mathbb{E}_w \left[\mathbb{E}_x \left[\phi^2(w^\top f(x)) \right] - \mathbb{E}_{xy} \left[\phi(w^\top f(x)) \phi(w^\top f(y)) \right] \right] - \xi \right| = O\left(N_{l-1}^{-1/2} \right), \tag{C.110}$$

with

$$\begin{aligned}\xi :=&\mathbb{E}_x\mathbb{E}_w\left[\phi^2\left(w^\top f(x)\frac{\tilde{c}}{\tilde{c}(x)}\right)\right]-\mathbb{E}_{xy}\mathbb{E}_w\left[\phi\left(w^\top f(x)\frac{\tilde{c}}{\tilde{c}(x)}\right)\phi\left(w^\top f(y)\frac{\tilde{c}}{\tilde{c}(y)}\right)\right]\\ =&\mathbb{E}_{\rho_1}\left[\tilde{\phi}^2(\rho_1)\right]-\mathbb{E}_{xy}\left[\mathbb{E}_{\rho_1\rho_2}\left[\tilde{\phi}(\rho_1)\tilde{\phi}(\rho_2)\right]\right],\end{aligned} \tag{C.111}$$

where we have set $\tilde{\phi}(t)=\phi(\tilde{c}t)$. Hence, in order to obtain that the quantity in (C.107) is $\Theta(1)$, it suffices to prove that $\xi=\Theta(1)$.

As ϕ is Lipschitz and $\tilde{c}$ is $\Theta(1)$, $\tilde{\phi}$ is also Lipschitz, which readily implies that $\xi=O(1)$. We now prove that $\xi=\Omega(1)$. By exploiting the Hermite expansion of $\tilde{\phi}$, we have that

$$\xi=\sum_{i=0}^{\infty}\mu_i^2\left(1-\mathbb{E}_{xy}\left[\left(\frac{f(x)^\top f(y)}{\|f(x)\|\,\|f(y)\|}\right)^i\right]\right), \tag{C.112}$$

where μ_i is the i-th Hermite coefficient of $\tilde{\phi}$. Note that, since we conditioned on $\tilde{c}=\Theta(1)$, these coefficients are numerical constants. As ϕ (and therefore $\tilde{\phi}$) is non constant, there exist $j>0$ such that $\mu_j\neq 0$. Furthermore, we have that the sum in (C.112) contains only positive terms, as $|f(x)^\top f(y)|\leq\|f(x)\|\cdot\|f(y)\|$ by Cauchy-Schwarz. Therefore, in order to show that $\xi=\Omega(1)$, it suffices to prove that, for all $j>0$,

$$\mathbb{E}_{xy}\left[\left(\frac{\left|f(x)^\top f(y)\right|}{\|f(x)\|\,\|f(y)\|}\right)^j\right]\leq C_0<1, \tag{C.113}$$

where C_0 is an absolute constant strictly smaller than 1. Furthermore, (C.113) is implied by the following:

$$\mathbb{P}_{xy}\left(\frac{\left|f(x)^\top f(y)\right|}{\|f(x)\|\,\|f(y)\|}\leq C_1\right)\geq c_1, \tag{C.114}$$

where $C_1<1$ and $c_1>0$ are numerical constants.

By writing $f(x)$ as $\tilde{f}(x)+f$ and $f(y)$ as $\tilde{f}(y)+f$, we have

$$\begin{aligned}\frac{\left|f(x)^\top f(y)\right|}{\|f(x)\|\,\|f(y)\|}&=\frac{\left|(\tilde{f}(x)+f)^\top(\tilde{f}(y)+f)\right|}{\left\|\tilde{f}(x)+f\right\|\left\|\tilde{f}(y)+f\right\|}\\ &\leq\frac{\left|\tilde{f}(x)^\top(\tilde{f}(y)+f)\right|+\left|f^\top\tilde{f}(y)\right|+\|f\|^2}{\min_{z\in\{x,y\}}\left\|\tilde{f}(z)+f\right\|^2}\\ &\leq\frac{\left|\tilde{f}(x)^\top f(y)\right|+\left|f^\top\tilde{f}(y)\right|+\|f\|^2}{\min_{z\in\{x,y\}}\left(\left\|\tilde{f}(z)\right\|^2-2\left|f^\top\tilde{f}(z)\right|\right)+\|f\|^2}.\end{aligned} \tag{C.115}$$

Let us provide bounds on the various terms appearing in (C.115):

(i) Part (d) of the conditioning (cf. (C.100)) gives that

$$\mathbb{P}_{xy}\left(\min_{z\in\{x,y\}}\left\|\tilde{f}(z)\right\|_2^2\geq cN_{l-1}\right)\geq\frac{1}{4}, \tag{C.116}$$

for some numerical constant $c>0$.

(ii) Part (b) of the conditioning gives that

$$\|f\|_2^2 \leq C' N_{l-1}. \tag{C.117}$$

(iii) Part (a) of the conditioning gives that $\|f_{l-1}(x)\|_{\text{Lip}} = O(1)$, and part (b) of the conditioning gives that $\mathbb{E}_y[\|f_{l-1}(y)\|] = \Theta(\sqrt{N_{l-1}})$. Hence, as $y \sim P_X$, Assumption 2.2 implies that

$$\|f_{l-1}(y)\| = \Theta(\sqrt{N_{l-1}}), \tag{C.118}$$

with probability at least $1 - 2\exp(-cN_{l-1})$ over y, where c is a numerical constant. Furthermore, by recalling that $\mathbb{E}_x[\tilde{f}(x)] = 0$ and using again Assumption 2.2, we have that, for any fixed vector u,

$$\mathbb{P}_x(|\tilde{f}(x)^\top u| > t) \leq 2e^{-c_0 t^2/\|u\|_2^2}, \tag{C.119}$$

where c_0 is another numerical constant. Since x and y are independent, (C.119) implies that

$$\mathbb{P}_x(|\tilde{f}(x)^\top f(y)| > N_{l-1}^{3/4}) \leq 2e^{-c_0 N_{l-1}^{3/2}/\|f(y)\|_2^2} \leq 2e^{-c_2 N_{l-1}^{1/2}}, \tag{C.120}$$

where the first inequality holds for every y, and the second inequality holds with probability at least $1 - 2\exp(-cN_{l-1})$ over y by (C.118). As a result, we have

$$\mathbb{P}_{xy}(|\tilde{f}(x)^\top f(y)| > N_{l-1}^{3/4}) \leq 2e^{-c_2\sqrt{N_{l-1}}} + 2e^{-cN_{l-1}} \leq 4e^{-c_3\sqrt{N_{l-1}}}. \tag{C.121}$$

(iv) By setting $t = N_{l-1}^{3/4}$ and $u = f$ into (C.119), we obtain

$$\mathbb{P}_y(|\tilde{f}(y)^\top f| > N_{l-1}^{3/4}) \leq 2e^{-c_4\sqrt{N_{l-1}}}, \tag{C.122}$$

where c_4 is a numerical constant and we have also used (C.117).

(v) Finally, as x and y are independent, (C.122) implies that

$$\mathbb{P}_{xy}(\max_{z\in\{x,y\}} |\tilde{f}(z)^\top f| > N_{l-1}^{3/4}) \leq 4e^{-c_4\sqrt{N_{l-1}}}. \tag{C.123}$$

By plugging into (C.115) the bounds (C.116), (C.121), (C.122) and (C.123), we obtain that

$$\mathbb{P}_{x,y}\left(\frac{\left|f(x)^\top f(y)\right|}{\|f(x)\|\,\|f(y)\|} \leq \frac{2N_{l-1}^{3/4} + \|f\|_2^2}{c\,N_{l-1} - 2N_{l-1}^{3/4} + \|f\|_2^2}\right) \geq 1/4 - 10e^{-c_5\sqrt{N_{l-1}}}. \tag{C.124}$$

By using also (C.117), we have that (C.114) readily follows from (C.124). From (C.114), we have that $\xi = \Theta(1)$. Hence, the quantity in (C.107) is $\Theta(1)$ and in particular it is lower bounded by a numerical constant, call it C_0.

By setting $t = C_0/2$ in (C.106), we conclude that

$$\mathbb{E}_x\left[\|f_l(x) - \mathbb{E}_x[f_l(x)]\|_2^2\right] = \Theta(N_l), \tag{C.125}$$

with probability at least $1 - 2\exp(-cN_{l-1})$ over W_l, where c is an absolute constant. By taking into account the conditioning made at the beginning of the proof over the space $(W_k)_{k=1}^{l-1}$, we obtain that (C.125) holds with probability at least $1 - 5C'\exp(-cN_{n-1}) - 2\exp(-cN_{n-1}) \geq$

$1 - C\exp(-cN_{n-1})$ over $(W_k)_{k=1}^l$, where C is a numerical constant, which concludes the proof of (C.96).

Finally, we prove (C.97) and (C.98), again for layer l. By Lemma C.2, we have that $\|f_l(x)\|_{\mathrm{Lip}} = O(1)$, with probability at least $1 - C'\exp(-cN_{L-1})$ over $(W_k)_{k=1}^l$. By conditioning on this event, we also have that $\|\|f_l(x) - \mathbb{E}_x[f_l(x)]\|_2\|_{\mathrm{Lip}} = O(1)$. Furthermore, we condition on a realization of $(W_k)_{k=1}^l$ such that (C.96) holds.

To obtain (C.97), we apply Jensen's inequality and Lemma C.13, which give that

$$\mathbb{E}_x\left[\|f_l(x) - \mathbb{E}_x[f_l(x)]\|_2\right] = \Theta(\sqrt{N_l}), \tag{C.126}$$

with probability at least $1 - C'\exp(-cN_{L-1}) - 5C'\exp(-cN_{n-1}) - 2\exp(-cN_{n-1}) \geq 1 - C\exp(-cN_{n-1})$ over $(W_k)_{k=1}^l$.

To obtain (C.98), we condition on a realization of $(W_k)_{k=1}^l$ such that $\|\|f_l(x) - \mathbb{E}_x[f_l(x)]\|_2\|_{\mathrm{Lip}} = O(1)$ and (C.97) holds. Then, by Assumption 2.2, we have that

$$\begin{aligned}\mathbb{P}_x\left(\left|\|f_l(x) - \mathbb{E}_x[f_l(x)]\|_2 - \mathbb{E}_x\left[\|f_l(x) - \mathbb{E}_x[f_l(x)]\|_2\right]\right| > \mathbb{E}_x\left[\|f_l(x) - \mathbb{E}_x[f_l(x)]\|_2\right]/2\right)\\ \leq 2\exp(-c_1 N_l) \leq 2\exp(-cN_{L-1}),\end{aligned} \tag{C.127}$$

where c is a numerical constant. This gives that

$$\|f_l(x) - \mathbb{E}_x[f_l(x)]\| = \Theta(\sqrt{N_l}), \tag{C.128}$$

with probability at least $1 - 6C'\exp(-cN_{n-1}) - 2\exp(-cN_{n-1}) - 2\exp(-cN_{n-1}) \geq 1 - C\exp(-cN_{n-1})$ over x and $(W_k)_{k=1}^l$, which concludes the proof. □

Lemma C.20 (L2 norms of centered backpropagation)**.** *Let $x \sim P_X$. Then, we have*

$$\|D_L\phi'(g_{L-1}(x)) - \mathbb{E}_x\left[D_L\phi'(g_{L-1}(x))\right]\|_2 = \Theta(\sqrt{N_{L-1}}), \tag{C.129}$$

with probability at least $1 - 10\exp(-c\log^2 N_{L-1}) - C\exp(-cN_{L-1})$ over x and $(W_k)_{k=1}^L$ over $(W_k)_{k=1}^l$ and x.

Proof. An application of Lemma C.19 for $l = L - 2$ gives that

$$\|f_{L-2}(x) - \mathbb{E}_x\left[f_{L-2}(x)\right]\|_2 = \Theta(\sqrt{N_{L-2}}). \tag{C.130}$$

with probability at least $1 - C'\exp(-cN_{L-1})$ over $(W_k)_{k=1}^{L-2}$ and x.

To ease the notation, we use the shorthands $f(x) := f_{L-2}(x)$, $f = \mathbb{E}_x[f(x)]$, $\tilde{f}(x) = f(x) - f$, $W := W_{L-1}$ and $w_i = W_{:i}$. We also define $\tilde{c}(x) = \beta_l\|f(x)\|/\sqrt{N_{l-1}}$ and $\tilde{c} = \mathbb{E}_x[\tilde{c}(x)]$.

As in Lemma C.19, we condition on the 3 events (a)-(c), which jointly happen with probability at least $1 - 4C'\exp(-cN_{L-1})$ over $(W_k)_{k=1}^{L-2}$. Note that, to condition on the event (c), we use (C.130).

Now, we can write

$$\begin{aligned}
&\mathbb{E}_x\left[\|D_L\phi'(g_{L-1}(x)) - \mathbb{E}_x\left[D_L\phi'(g_{L-1}(x))\right]\|_2^2\right] \\
&= \mathbb{E}_x\left[\left\|D_L\phi'(W^\top f_{L-2}(x)) - \mathbb{E}_x\left[D_L\phi'(W^\top f_{L-2}(x))\right]\right\|_2^2\right] \\
&= N_{L-1}\left(\frac{1}{N_{L-1}}\sum_{i=1}^{N_{L-1}}(D_L)_{ii}^2\mathbb{E}_x\left[\left(\phi'(w_i^\top f(x)) - \mathbb{E}_x\left[\phi'(w_i^\top f(x))\right]\right)^2\right]\right) \\
&= N_{L-1}\left(\frac{1}{N_{L-1}}\sum_{i=1}^{N_{L-1}}(D_L)_{ii}^2\mathbb{E}_{xw_1}\left[\left(\phi'(w_1^\top f(x)) - \mathbb{E}_x\left[\phi'(w_1^\top f(x))\right]\right)^2\right] + \frac{1}{N_{L-1}}\sum_{i=1}^{N_{L-1}} Z_i\right),
\end{aligned} \tag{C.131}$$

where we use that the w_i-s are identically distributed and we have defined the independent, mean-0 random variables

$$\begin{aligned}
Z_i =&(D_L)_{ii}^2\mathbb{E}_x\left[\left(\phi'(w_i^\top f(x)) - \mathbb{E}_x\left[\phi'(w_i^\top f(x))\right]\right)^2\right] \\
&- (D_L)_{ii}^2\mathbb{E}_{w_1x}\left[\left(\phi'(w_1^\top f(x)) - \mathbb{E}_x\left[\phi'(w_1^\top f(x))\right]\right)^2\right].
\end{aligned} \tag{C.132}$$

Note that in the definition of Z_i the randomness comes only from w_i and $(D_L)_{ii}$, since we are conditioning on $(W_k)_{k=1}^{L-2}$.

If we fix the $(D_L)_{ii}$-s and follow the same argument in (C.103)-(C.105) (cf. the proof of Lemma C.19), we have

$$\|Z_i\|_{\psi_1} \leq C_0(D_L)_{ii}^2, \tag{C.133}$$

where C_0 is a numerical constant and we have used that ϕ' is Lipschitz. Let $\mathcal{E}_{\text{bad}}$ be the event s.t. $\max_i(D_L)_{ii}^2 > \log^2 N_{L-1}$. Then, by following the same argument as in Lemma C.3, we have that

$$\mathbb{P}(\mathcal{E}_{\text{bad}}) \leq 2\exp(-c\log^2 N_{L-1}). \tag{C.134}$$

Hence, by conditioning on $\mathcal{E}_{\text{bad}}^c$, we have that

$$\max_i \|Z_i\|_{\psi_1} \leq C_0\log^2 N_{L-1}. \tag{C.135}$$

By applying Bernstein inequality (cf. Corollary 2.8.3. in Vershynin [2018]), we get

$$\mathbb{P}\left(\left|\frac{1}{N_{L-1}}\sum_{i=1}^{N_{L-1}} Z_i\right| > \frac{1}{\sqrt[4]{N_{L-1}}}\,\middle|\,\mathcal{E}_{\text{bad}}^c\right) \leq 2\exp\left(-c\frac{\sqrt{N_{L-1}}}{\log^4 N_{L-1}}\right), \tag{C.136}$$

for some numerical constant c. By combining (C.134) and (C.136), we obtain that

$$\begin{aligned}
\mathbb{P}\left(\left|\frac{1}{N_{L-1}}\sum_{i=1}^{N_{L-1}} Z_i\right| > \frac{1}{\sqrt[4]{N_{L-1}}}\right) &\leq 2\exp\left(-c\frac{\sqrt{N_{L-1}}}{\log^4 N_{L-1}}\right) + 2\exp\left(-c\log^2 N_{L-1}\right) \\
&\leq 4\exp\left(-c\log^2 N_{L-1}\right),
\end{aligned} \tag{C.137}$$

where this probability is over W_{L-1} and W_L.

Let's now focus on the first term in the last line of (C.131). In particular, we have that

$$\xi = \mathbb{E}_{xw_1}\left[\left(\phi'(w_1^\top f(x)) - \mathbb{E}_x\left[\phi'(w_1^\top f(x))\right]\right)^2\right] = \Theta(1). \tag{C.138}$$

This can be proven by following the same argument in (C.108)-(C.124) (cf. the proof of Lemma C.19), as ϕ' is Lipschitz and non-constant.

Next, we re-write (C.131) as

$$\begin{aligned}
&\mathbb{E}_x\left[\|D_L\phi'(g_{L-1}(x)) - \mathbb{E}_x\left[D_L\phi'(g_{L-1}(x))\right]\|_2^2\right] \\
&= N_{L-1}\left(\frac{1}{N_{L-1}}\sum_{i=1}^{N_{L-1}}(D_L)_{ii}^2\mathbb{E}_{xw_1}\left[\left(\phi'(w_1^\top f(x)) - \mathbb{E}_x\left[\phi'(w_1^\top f(x))\right]\right)^2\right] + \frac{1}{N_{L-1}}\sum_{i=1}^{N_{L-1}} Z_i\right) \\
&= N_{L-1}\left(\xi\mathbb{E}_{W_L}\left[(D_L)_{11}^2\right] + \xi\frac{1}{N_{L-1}}\sum_{i=1}^{N_{L-1}}\tilde{Z}_i + \frac{1}{N_{L-1}}\sum_{i=1}^{N_{L-1}} Z_i\right) \\
&= N_{L-1}\left(\xi\beta_L^2 + \xi\frac{1}{N_{L-1}}\sum_{i=1}^{N_{L-1}}\tilde{Z}_i + \frac{1}{N_{L-1}}\sum_{i=1}^{N_{L-1}} Z_i\right),
\end{aligned} \tag{C.139}$$

where we have defined the independent, mean-0, sub-exponential random variables

$$\tilde{Z}_i = (D_L)_{ii}^2 - \mathbb{E}_{W_L}\left[(D_L)_{11}^2\right]. \tag{C.140}$$

Since the $(D_L)_{ii}$-s are standard Gaussian, we have

$$\left\|\tilde{Z}_i\right\|_{\psi_1} \leq \left\|(D_L)_{ii}^2\right\|_{\psi_1} = \|(D_L)_{ii}\|_{\psi_2}^2 = C_1. \tag{C.141}$$

Hence, another application of Bernstein inequality (cf. Corollary 2.8.3. in Vershynin [2018]) allows us to conclude that

$$\left|\frac{1}{N_{L-1}}\sum_{i=1}^{N_{L-1}}\tilde{Z}_i\right| = O\left(N_{L-1}^{-1/4}\right), \tag{C.142}$$

with probability at least $1 - 2\exp(-c\sqrt{N_{L-1}})$ over W_L.

Thus, by using (C.137) and (C.142), and taking into account the initial conditioning over $(W_k)_{k=1}^{L-2}$, we conclude that

$$\mathbb{E}_x\left[\|D_L\phi'(g_{L-1}(x)) - \mathbb{E}_x\left[D_L\phi'(g_{L-1}(x))\right]\|_2^2\right] = \Theta(N_{L-1}), \tag{C.143}$$

with probability at least $1 - 6\exp(-c\log^2 N_{L-1}) - 6C'\exp(-cN_{L-1})$ over $(W_k)_{k=1}^L$.

Proceeding in a similar fashion as in Lemma C.19, we apply Lemma C.4, which gives that $\|D_L\phi'(g_{L-1}(x))\|_{\text{Lip}} = O(\log N_{L-1})$, with probability at least $1 - 2\exp(-c\log^2 N_{L-1}) - C'\exp(-N_{L-1})$ over $(W_k)_{k=1}^L$. By conditioning on this event, we also have that $\|\|D_L\phi'(g_{L-1}(x)) - \mathbb{E}_x\left[D_L\phi'(g_{L-1}(x))\right]\|_2\|_{\text{Lip}} = O(\log N_{L-1})$. Furthermore, we condition on a realization of $(W_k)_{k=1}^L$ such that (C.143) holds.

We can now apply Jensen's inequality and Lemma C.13, to obtain that

$$\mathbb{E}_x\left[\|D_L\phi'(g_{L-1}(x)) - \mathbb{E}_x\left[D_L\phi'(g_{L-1}(x))\right]\|_2\right] = \Theta(\sqrt{N_l}), \tag{C.144}$$

with probability at least $1 - 8\exp(-c\log^2 N_{L-1}) - 7C'\exp(-cN_{L-1})$ over $(W_k)_{k=1}^L$.

Finally, we condition on a realization of $(W_k)_{k=1}^L$ such that $\|\|D_L\phi'(g_{L-1}(x)) - \mathbb{E}_x\left[D_L\phi'(g_{L-1}(x))\right]\|_2\|_{\text{Lip}} = O(\log N_{L-1})$ and (C.144) hold. Then, by

Assumption 2.2, we have that

$$\mathbb{P}_x\Big(\big|\|D_L\phi'(g_{L-1}(x)) - \mathbb{E}_x\left[D_L\phi'(g_{L-1}(x))\right]\|_2 - \mathbb{E}_x\left[\|D_L\phi'(g_{L-1}(x)) - \mathbb{E}_x\left[D_L\phi'(g_{L-1}(x))\right]\|_2\right]\big| \\ > \mathbb{E}_x\left[\|D_L\phi'(g_{L-1}(x)) - \mathbb{E}_x\left[D_L\phi'(g_{L-1}(x))\right]\|_2\right]/2\Big) \\ \leq 2\exp^(-cN_{L-1}). \tag{C.145}$$

This gives that

$$\|D_L\phi'(g_{L-1}(x)) - \mathbb{E}_x\left[D_L\phi'(g_{L-1}(x))\right]\|_2 = \Theta(\sqrt{N_{L-1}}), \tag{C.146}$$

with probability at least $1 - 10\exp(-c\log^2 N_{L-1}) - 8C'\exp(-cN_{L-1})$ over x and $(W_k)_{k=1}^L$. This concludes the proof. □

C.3 Proofs for Part 1: Centering

C.3.1 Step (a): Centering Features and Backpropagation

Lemma C.21 (Centering features and backpropagation)**.** *Consider the setting of Theorem 2.6, let $F_{L-2} \in \mathbb{R}^{n\times N_{L-2}}$ be the feature matrix at layer $L-2$, and let B_{L-1} contain the backpropagation terms from layer $L-1$, i.e. $(B_{L-1})_{i:} = D_L\phi'(g_{L-1}(x_i))$. Let $J_{L-2}J_{L-2}^\top = F_{L-2}F_{L-2}^\top \circ B_{L-1}B_{L-1}^\top$ and $\tilde{J}_{FB}\tilde{J}_{FB}^\top = \tilde{F}_{L-2}\tilde{F}_{L-2}^\top \circ \tilde{B}_{L-1}\tilde{B}_{L-1}^\top$, where $\tilde{F}_{L-2} = F_{L-2} - \mathbb{E}_X[F_{L-2}]$ and $\tilde{B}_{L-1} = B_{L-1} - \mathbb{E}_X[B_{L-1}]$. Then, we have that*

$$\lambda_{\min}\left(J_{L-2}J_{L-2}^\top\right) \geq \lambda_{\min}\left(\tilde{J}_{FB}\tilde{J}_{FB}^\top\right) - o(N_{L-1}N_{L-2}), \tag{C.147}$$

with probability at least $1 - C\exp(-cN_{L-1}) - 4\exp(-c\log^2 n) - 8\exp(-c\log^2 N_{L-1})$ over $(W_k)_{k=1}^L$ and $(x_i)_{i=1}^n \sim_{\text{i.i.d.}} P_X$, where c and C are numerical constants.

Proof. By Lemma C.2 and Lemma C.4, we have that $\|f_{L-2}\|_{\text{Lip}} = O(1)$ and $\|D_L\phi'(g_{L-1}(x))\|_{\text{Lip}} = O(\log N_{L-1})$ with probability $1 - C\exp(-N_{L-1}) - 2\exp(-c\log^2 N_{L-1})$ over $(W_k)_{k=1}^L$. We will condition on these events for the rest of the proof.

Let's define $\tilde{J}_F\tilde{J}_F^\top = \tilde{F}_{L-2}\tilde{F}_{L-2}^\top \circ B_{L-1}B_{L-1}^\top$. We can now re-write the quantity $J_{L-2}J_{L-2}^\top$ as follows:

$$\begin{aligned} J_{L-2}J_{L-2}^\top &= \tilde{J}_F\tilde{J}_F^\top + \mathbb{E}[F_{L-2}]\mathbb{E}[F_{L-2}]^\top \circ B_{L-1}B_{L-1}^\top \\ &\quad + \Big((F_{L-2} - \mathbb{E}[F_{L-2}])\mathbb{E}[F_{L-2}]^\top + \mathbb{E}[F_{L-2}](F_{L-2} - \mathbb{E}[F_{L-2}])^\top\Big) \circ B_{L-1}B_{L-1}^\top \\ &= \tilde{J}_F\tilde{J}_F^\top + \|\nu\|_2^2 B_{L-1}B_{L-1}^\top + \left(\Lambda\mathbf{1}\mathbf{1}^\top + \mathbf{1}\mathbf{1}^\top\Lambda\right) \circ B_{L-1}B_{L-1}^\top \\ &= \tilde{J}_F\tilde{J}_F^\top + \left(\|\nu\|_2\mathbf{1} + \frac{\Lambda\mathbf{1}}{\|\nu\|_2}\right)\left(\|\nu\|_2\mathbf{1} + \frac{\Lambda\mathbf{1}}{\|\nu\|_2}\right)^\top \circ B_{L-1}B_{L-1}^\top \\ &\quad - \frac{\Lambda\mathbf{1}\mathbf{1}^\top\Lambda}{\|\nu\|_2^2} \circ B_{L-1}B_{L-1}^\top \\ &\succeq \tilde{J}_F\tilde{J}_F^\top - \frac{\Lambda\mathbf{1}\mathbf{1}^\top\Lambda}{\|\nu\|_2^2} \circ B_{L-1}B_{L-1}^\top, \end{aligned} \tag{C.148}$$

where $\nu = \mathbb{E}_{x_i}[(F_{L-2})_{i:}] \in \mathbb{R}^{N_{L-2}}$ (independent on i, since the x_i-s are i.i.d.), Λ is a diagonal matrix such that $\Lambda_{ii} = \nu^\top(F_{L-2})_{i:} - \|\nu\|_2^2 =: \mu(x_i)$, and $\mathbf{1} \in \mathbb{R}^n$ is a vector full of ones. The

last step is justified since the Hadamard product of PSD matrices is PSD by the Schur product theorem. Notice that we are assuming $\|\nu\|_2 \neq 0$. In fact, if $\|\nu\|_2 = 0$, then we immediately have that $J = \tilde{J}_F$.

Expanding in an analogous way the term $\tilde{J}_F \tilde{J}_F^\top$, we get

$$\begin{aligned} J_{L-2}J_{L-2}^\top &\succeq \tilde{J}_F\tilde{J}_F^\top - \frac{\Lambda \mathbf{1}\mathbf{1}^\top \Lambda}{\|\nu\|_2^2} \circ B_{L-1}B_{L-1}^\top \\ &= \tilde{J}_{FB}\tilde{J}_{FB}^\top + \left(\|\eta\|_2 \mathbf{1} + \frac{\Gamma \mathbf{1}}{\|\eta\|_2}\right)\left(\|\eta\|_2 \mathbf{1} + \frac{\Gamma \mathbf{1}}{\|\eta\|_2}\right)^\top \circ \tilde{F}_{L-2}\tilde{F}_{L-2}^\top \\ &\quad - \frac{\Gamma \mathbf{1}\mathbf{1}^\top \Gamma}{\|\eta\|_2^2} \circ \tilde{F}_{L-2}\tilde{F}_{L-2}^\top - \frac{\Lambda \mathbf{1}\mathbf{1}^\top \Lambda}{\|\nu\|_2^2} \circ B_{L-1}B_{L-1}^\top \\ &\succeq \tilde{J}_{FB}\tilde{J}_{FB}^\top + \left(\|\eta\|_2 \mathbf{1} + \frac{\Gamma \mathbf{1}}{\|\eta\|_2}\right)\left(\|\eta\|_2 \mathbf{1} + \frac{\Gamma \mathbf{1}}{\|\eta\|_2}\right)^\top \circ \tilde{F}_{L-2}\left(\frac{\nu\nu^\top}{\|\nu\|_2^2}\right)\tilde{F}_{L-2}^\top \\ &\quad - \frac{\Gamma \mathbf{1}\mathbf{1}^\top \Gamma}{\|\eta\|_2^2} \circ \tilde{F}_{L-2}\tilde{F}_{L-2}^\top - \frac{\Lambda \mathbf{1}\mathbf{1}^\top \Lambda}{\|\nu\|_2^2} \circ B_{L-1}B_{L-1}^\top \end{aligned} \tag{C.149}$$

where $\eta = \mathbb{E}_{x_i}[(B_{L-1})_{i:}] \in \mathbb{R}^{N_{L-1}}$ (independent on i, since the x_i-s are i.i.d.), Γ is a diagonal matrix such that $\Gamma_{ii} = \eta^\top (B_{L-1})_{i:} - \|\eta\|_2^2 =: \zeta(x_i)$. The last step is justified by the fact that the following matrix is PSD

$$\left(\|\eta\|_2 \mathbf{1} + \frac{\Gamma \mathbf{1}}{\|\eta\|_2}\right)\left(\|\eta\|_2 \mathbf{1} + \frac{\Gamma \mathbf{1}}{\|\eta\|_2}\right)^\top \circ \tilde{F}_{L-2}\left(I - \frac{\nu\nu^\top}{\|\nu\|_2^2}\right)\tilde{F}_{L-2}^\top,$$

since it is the Hadamard product of two PSD matrices. Notice that we are assuming $\|\eta\|_2 \neq 0$. In fact, if $\|\eta\|_2 = 0$, then we immediately have that $\tilde{J}_F = \tilde{J}_{FB}$.

Taking into account the following relations

$$\Lambda \mathbf{1} = \tilde{F}_{L-2}\nu, \qquad \Gamma \mathbf{1} = \tilde{B}_{L-1}\eta, \qquad \mathbb{E}_X[B_{L-1}] = \mathbf{1}\eta^\top, \tag{C.150}$$

we can simplify the second and the fourth term of the RHS of equation (C.149) as follows

$$\begin{aligned} &\left(\|\eta\|_2 \mathbf{1} + \frac{\Gamma \mathbf{1}}{\|\eta\|_2}\right)\left(\|\eta\|_2 \mathbf{1} + \frac{\Gamma \mathbf{1}}{\|\eta\|_2}\right)^\top \circ \tilde{F}_{L-2}\left(\frac{\nu\nu^\top}{\|\nu\|_2^2}\right)\tilde{F}_{L-2}^\top - B_{L-1}B_{L-1}^\top \circ \frac{\Lambda \mathbf{1}\mathbf{1}^\top \Lambda}{\|\nu\|_2^2} \\ &= \left(\|\eta\|_2 \mathbf{1} + \frac{\Gamma \mathbf{1}}{\|\eta\|_2}\right)\left(\|\eta\|_2 \mathbf{1} + \frac{\Gamma \mathbf{1}}{\|\eta\|_2}\right)^\top \circ \frac{\Lambda \mathbf{1}\mathbf{1}^\top \Lambda}{\|\nu\|_2^2} \\ &\quad - \left(\mathbf{1}\eta^\top + \tilde{B}_{L-1}\right)\left(\mathbf{1}\eta^\top + \tilde{B}_{L-1}\right)^\top \circ \frac{\Lambda \mathbf{1}\mathbf{1}^\top \Lambda}{\|\nu\|_2^2} \\ &= \left(\frac{\Gamma \mathbf{1}\mathbf{1}^\top \Gamma}{\|\eta\|_2^2} - \tilde{B}_{L-1}\tilde{B}_{L-1}^\top\right) \circ \frac{\Lambda \mathbf{1}\mathbf{1}^\top \Lambda}{\|\nu\|_2^2} \succeq -\tilde{B}_{L-1}\tilde{B}_{L-1}^\top \circ \frac{\Lambda \mathbf{1}\mathbf{1}^\top \Lambda}{\|\nu\|_2^2}. \end{aligned} \tag{C.151}$$

Merging this last relation with (C.149) we get

$$\begin{aligned} J_{L-2}J_{L-2}^\top &\succeq \tilde{J}_{FB}\tilde{J}_{FB}^\top - \frac{\Gamma \mathbf{1}\mathbf{1}^\top \Gamma}{\|\eta\|_2^2} \circ \tilde{F}_{L-2}\tilde{F}_{L-2}^\top - \frac{\Lambda \mathbf{1}\mathbf{1}^\top \Lambda}{\|\nu\|_2^2} \circ \tilde{B}_{L-1}\tilde{B}_{L-1}^\top \\ &= \tilde{J}_{FB}\tilde{J}_{FB}^\top - \left(\frac{\Gamma}{\|\eta\|_2}\tilde{F}_{L-2}\right)\left(\frac{\Gamma}{\|\eta\|_2}\tilde{F}_{L-2}\right)^\top - \left(\frac{\Lambda}{\|\nu\|_2}\tilde{B}_{L-1}\right)\left(\frac{\Lambda}{\|\nu\|_2}\tilde{B}_{L-1}\right)^\top. \end{aligned} \tag{C.152}$$

Note that $\|\mu\|_{\text{Lip}} \leq \|f_{L-2}\|_{\text{Lip}} \|\nu\|_2$, and that $\mathbb{E}_{x_i}[\mu(x_i)] = 0$ for all $i \in [n]$. Thus, by using Assumption 2.2 on x_i and exploiting the initial conditioning on the weights, we have that

$$\mathbb{P}(|\mu(x_i)|/\|\nu\|_2 > t) < 2\exp(-ct^2), \tag{C.153}$$

where the probability is intended over $x_i \sim P_X$. Thus, following the same argument of Lemma C.3, the last relation implies that

$$\|\Lambda/\|\nu\|_2\|_{\text{op}} = O(\log n), \tag{C.154}$$

with probability at least $1 - 2\exp(-c\log^2 n)$, where c is a numerical constant, and the probability is intended over $\{x_i\}_{i=1}^n$. This implies that

$$\begin{aligned}
&\left\| \left(\frac{\Lambda}{\|\nu\|_2} \tilde{B}_{L-1} \right) \left(\frac{\Lambda}{\|\nu\|_2} \tilde{B}_{L-1} \right)^\top \right\|_{\text{op}} \\
&\qquad \leq \left\| \frac{\Lambda}{\|\nu\|_2} \right\|_{\text{op}}^2 \left\| \tilde{B}_{L-1} \tilde{B}_{L-1}^\top \right\|_{\text{op}} = O\left((n + N_{L-1}) \cdot \log^2 n \cdot \log^2 N_{L-1} \right) = o(N_{L-2} N_{L-1}),
\end{aligned} \tag{C.155}$$

where the second equality is justified by Lemma C.9, and the last by Lemma C.1. This result holds with probability $1 - C\exp(-cN_{L-1}) - 2\exp(-c\log^2 n) - 4\exp(-c\log^2 N_{L-1})$ over $(W_k)_{k=1}^L$ and $(x_i)_{i=1}^n$.

Note that $\|\zeta\|_{\text{Lip}} \leq \|D_L \phi'(g_{L-1}(x))\|_{\text{Lip}} \|\eta\|_2$, and that $\mathbb{E}_{x_i}[\zeta(x_i)] = 0$ for all $i \in [n]$. Thus, by using Assumption 2.2 on x_i and exploiting the initial conditioning on the weights, we have that

$$\mathbb{P}(|\zeta(x_i)|/\|\eta\|_2 > t \cdot \log N_{L-1}) < 2\exp(-ct^2), \tag{C.156}$$

where the probability is intended over $x_i \sim P_X$. Thus, following the same argument of Lemma C.3, the last relation implies that

$$\|\Gamma/\|\eta\|_2\|_{\text{op}} = O(\log n \cdot \log N_{L-1}), \tag{C.157}$$

with probability at least $1 - 2\exp(-c\log^2 n)$, where c is a numerical constant, and the probability is intended over $\{x_i\}_{i=1}^n$. This implies that

$$\begin{aligned}
&\left\| \left(\frac{\Gamma}{\|\eta\|_2} \tilde{F}_{L-2} \right) \left(\frac{\Gamma}{\|\eta\|_2} \tilde{F}_{L-2} \right)^\top \right\|_{\text{op}} \\
&\qquad \leq \left\| \frac{\Gamma}{\|\eta\|_2} \right\|_{\text{op}}^2 \left\| \tilde{F}_{L-2} \tilde{F}_{L-2}^\top \right\|_{\text{op}} = O\left((n + N_{L-2}) \cdot \log^2 n \cdot \log^2 N_{L-1} \right) = o(N_{L-2} N_{L-1}),
\end{aligned} \tag{C.158}$$

where the second equality is justified by Lemma C.8, and the last by Lemma C.1. This result holds with probability $1 - C\exp(-cN_{L-1}) - 2\exp(-c\log^2 n) - 4\exp(-c\log^2 N_{L-1})$ over $(W_k)_{k=1}^L$ and $(x_i)_{i=1}^n$.

By merging (C.155) and (C.158) with (C.152), we readily obtain the desired result. □

C.3.2 Step (b): Centering everything

Lemma C.22 (Centering everything)**.** *Consider the setting of Theorem 2.6, let $F_{L-2} \in \mathbb{R}^{n \times N_{L-2}}$ be the feature matrix at layer $L-2$, and let B_{L-1} contain backpropagation terms*

from layer $L-1$, i.e. $(B_{L-1})_{i:} = D_L\phi'(g_{L-1}(x_i))$. Let $\tilde{J}_{FB}\tilde{J}_{FB}^\top = \tilde{F}_{L-2}\tilde{F}_{L-2}^\top \circ \tilde{B}_{L-1}\tilde{B}_{L-1}^\top$ and $\tilde{J}_{L-2}\tilde{J}_{L-2}^\top = \tilde{F}_{L-2}\tilde{F}_{L-2}^\top \circ \tilde{B}_{L-1}\tilde{B}_{L-1}^\top - \mathbb{E}_X[\tilde{F}_{L-2}\tilde{F}_{L-2}^\top \circ \tilde{B}_{L-1}\tilde{B}_{L-1}^\top]$, where $\tilde{F}_{L-2} = F_{L-2} - \mathbb{E}_X[F_{L-2}]$ and $\tilde{B}_{L-1} = B_{L-1} - \mathbb{E}_X[B_{L-1}]$. Then, we have that

$$\lambda_{\min}\left(\tilde{J}_{FB}\tilde{J}_{FB}^\top\right) \geq \lambda_{\min}\left(\tilde{J}_{L-2}\tilde{J}_{L-2}^\top\right) - o(N_{L-1}N_{L-2}), \tag{C.159}$$

with probability at least $1 - C\exp(-N_{L-1}) - 2\exp(-c\log^2 N_{L-1}) - 2\exp(-c\log^2 n)$ over $(W_k)_{k=1}^L$ and $(x_i)_{i=1}^n$.

Proof. Note that the i-th row of $\tilde{J}_{FB}$ is now in the form

$$(\tilde{J}_{FB})_{i:} = \tilde{f}_{L-2}(x_i) \otimes (D_L\tilde{\phi}'(g_{L-1}(x_i))), \tag{C.160}$$

where we recall that $\tilde{f}_{L-2}(x_i) = f_{L-2}(x_i) - \mathbb{E}_{x_i}[f_{L-2}(x_i)]$ and $\tilde{\phi}'(g_{L-1}(x_i)) = \phi'(g_{L-1}(x_i)) - \mathbb{E}[\phi'(g_{L-1}(x_i))]$. Furthermore, $(\tilde{J}_{L-2})_{i:} = (\tilde{J}_{FB})_{i:} - \mathbb{E}_{x_i}[(\tilde{J}_{FB})_{i:}]$. Then, by following similar passages as in (C.148), we have

$$\begin{aligned}
\tilde{J}_{FB}\tilde{J}_{FB}^\top &= \tilde{J}_{L-2}\tilde{J}_{L-2}^\top + \mathbb{E}[\tilde{J}_{FB}]\mathbb{E}[\tilde{J}_{FB}]^\top \\
&\quad + (\tilde{J}_{FB} - \mathbb{E}[\tilde{J}_{FB}])\mathbb{E}[\tilde{J}_{FB}]^\top + \mathbb{E}[\tilde{J}_{FB}](\tilde{J}_{FB} - \mathbb{E}[\tilde{J}_{FB}])^\top \\
&= \tilde{J}_{L-2}\tilde{J}_{L-2}^\top + \|A\|_F^2\,\mathbf{1}\mathbf{1}^\top + \Lambda\mathbf{1}\mathbf{1}^\top + \mathbf{1}\mathbf{1}^\top\Lambda \\
&= \tilde{J}_{L-2}\tilde{J}_{L-2}^\top + \left(\|A\|_F\,\mathbf{1} + \frac{\Lambda\mathbf{1}}{\|A\|_F}\right)\left(\|A\|_F\,\mathbf{1} + \frac{\Lambda\mathbf{1}}{\|A\|_F}\right)^\top - \frac{\Lambda\mathbf{1}\mathbf{1}^\top\Lambda}{\|A\|_F^2} \\
&\succeq \tilde{J}_{L-2}\tilde{J}_{L-2}^\top - \frac{\Lambda\mathbf{1}\mathbf{1}^\top\Lambda}{\|A\|_F^2},
\end{aligned} \tag{C.161}$$

where we have defined

$$A = \mathbb{E}_{x_i}\left[\tilde{f}_{L-2}(x_i)(D_L\tilde{\phi}'(g_{L-1}(x_i)))^\top\right], \tag{C.162}$$

which is independent on i (as the x_i-s are i.i.d.), and Λ is a diagonal matrix that contains in the i-th position

$$\Lambda_{ii} = \tilde{f}_{L-2}(x_i)^\top A(D_L\tilde{\phi}'(g_{L-1}(x_i))) - \mathbb{E}_{x_i}\left[\tilde{f}_{L-2}(x_i)^\top A(D_L\tilde{\phi}'(g_{L-1}(x_i)))\right]. \tag{C.163}$$

The last passage of (C.161) is true since we are subtracting a PSD matrix.

An application of Lemma C.2 gives that $\|f_{L-2}(x_i)\|_{\mathrm{Lip}}$ and $\|\phi'(g_{L-1}(x_i))\|_{\mathrm{Lip}}$ are upper bounded by a numerical constant both with probability at least $1 - C\exp(-N_{L-1})$ over $(W_k)_{k=1}^{L-1}$. Let us condition on this event on the probability space of $(W_k)_{k=1}^{L-1}$. Then, we can apply Lemma C.5 with $u(x) := \tilde{f}_{L-2}(x)$ and $v(x) := \tilde{\phi}'(g_{L-1}(x))$, which implies that

$$\|\Lambda_{ii}\|_{\psi_1} < C\,\|AD_L\|_F \leq C\,\|A\|_F\,\|D_L\|_{\mathsf{op}} = O\left(\log N_{L-1}\right)\|A\|_F. \tag{C.164}$$

In (C.164), C is a numerical constant and the last equality holds with probability at least $1 - 2\exp(-c\log^2 N_{L-1})$ over W_L by Lemma C.3. Thus,

$$\mathbb{P}(|\Lambda_{ii}|/\|A\|_F > t\cdot\log N_{L-1}) < 2\exp(-ct), \tag{C.165}$$

where the probability is intended over $x_i \sim P_X$ and c is a numerical constant. Thus, following the same argument of Lemma C.3, the last relation implies that

$$\|\Lambda/\|A\|_F\|_{\mathsf{op}} = O\left(\log^2 n\cdot\log N_{L-1}\right), \tag{C.166}$$

with probability at least $1 - 2\exp(-c\log^2 n)$, where c is a numerical constant, and the probability is intended over $\{x_i\}_{i=1}^n$.

Thus, with probability $1 - C\exp(-N_{L-1}) - 2\exp(-c\log^2 N_{L-1}) - 2\exp(-c\log^2 n)$ over $(W_k)_{k=1}^L$ and $(x_i)_{i=1}^n$, we have

$$\left\| \frac{\Lambda \mathbf{1}\mathbf{1}^\top \Lambda}{\|A\|_F^2} \right\|_{\mathsf{op}} \leq \|\mathbf{1}\|_{\mathsf{op}}^2 \left\| \frac{\Lambda}{\|A\|_F} \right\|_{\mathsf{op}}^2 = O\left(n \cdot \log^4 n \cdot \log^2 N_{L-1}\right) = o(N_{L-2}N_{L-1}), \quad \text{(C.167)}$$

where the last equality is a consequence of Lemma C.1.

The desired result follows from merging (C.167) with (C.161). □

C.4 Proofs for Part 2: Bounding the Centered Jacobian

C.4.1 L2 and Sub-Exponential Norms of Centered Jacobian

We start by providing an upper bound on the quantity $\mathbb{E}_x\left[\tilde{f}_{L-2}(x)(D_L\tilde{\phi}'(g_{L-1}(x)))^\top\right]$. This preliminary result will be useful when bounding the ℓ_2 norm of the rows of the centered Jacobian.

Lemma C.23. *Consider the setting of Theorem 2.6, let $x \sim P_X$, and let A be defined as*

$$A = \mathbb{E}_x\left[\tilde{f}_{L-2}(x)(D_L\tilde{\phi}'(g_{L-1}(x)))^\top\right]. \quad \text{(C.168)}$$

Then, we have

$$\|A\|_F = O\left(\sqrt{N_{L-1}}\log(N_{L-1})\right), \quad \text{(C.169)}$$

with probability at least $1 - 2\exp(-c\log^2 N_{L-1}) - C\exp(-cN_{L-1})$ over $(W_k)_{k=1}^L$, where c is an absolute constant.

Proof. We condition on $\|f_{L-2}(x)\|_{\mathrm{Lip}} = O(1)$ and on $\|\phi'(g_{L-1}(x))\|_{\mathrm{Lip}} = O(1)$. By Lemma C.2, these two conditions hold with probability at least $1 - C'\exp(-cN_{L-1})$ over $(W_k)_{k=1}^{L-1}$. Then, as P_X satisfies Assumption 2.2, we have that $\left\|\tilde{f}_{L-2}(x)\right\|_{\psi_2} = O(1)$ and $\left\|\tilde{\phi}'(g_{L-1}(x))\right\|_{\psi_2} = O(1)$. Hence, an application of Lemma C.6 gives that

$$\left\|\mathbb{E}_x\left[\tilde{f}_{L-2}(x)\tilde{\phi}'(g_{L-1}(x))^\top\right]\right\|_{\mathsf{op}} \leq C_1, \quad \text{(C.170)}$$

where C_1 is a numerical constant.

The following chain of inequalities holds:

$$\begin{aligned}
\|A\|_F &\leq \left\|\mathbb{E}_x\left[\tilde{f}_{L-2}(x)\tilde{\phi}'(g_{L-1}(x))^\top\right]\right\|_F \|D_L\|_{\mathsf{op}} \\
&\leq \sqrt{N_{L-1}}\left\|\mathbb{E}_x\left[\tilde{f}_{L-2}(x)\tilde{\phi}'(g_{L-1}(x))^\top\right]\right\|_{\mathsf{op}} \|D_L\|_{\mathsf{op}} \\
&\leq C_1\sqrt{N_{L-1}}\,\|D_L\|_{\mathsf{op}} \\
&= O\left(\sqrt{N_{L-1}}\log(N_{L-1})\right),
\end{aligned} \quad \text{(C.171)}$$

where the third line uses (C.170), and the last holds with probability $1 - 2\exp(-c\log^2 N_{L-1})$ over W_L, because of Lemma C.3. Taking into account the initial conditioning, we get the desired result. □

The next two results provide bounds on the ℓ_2 norm and on the sub-exponential ψ_1 norm of the rows of $\tilde{J}_{L-2}$, respectively.

Lemma C.24 (L2 norm of rows of centered Jacobian)**.** *Consider the setting of Theorem 2.6, let $x \sim P_X$, and let $\tilde{J}_x$ be defined as*

$$\tilde{J}_x = \tilde{f}_{L-2}(x) \otimes (D_L \tilde{\phi}'(g_{L-1}(x))) - \mathbb{E}_x \left[\tilde{f}_{L-2}(x) \otimes D_L \tilde{\phi}'(g_{L-1}(x)) \right]. \tag{C.172}$$

Then, we have

$$\left\| \tilde{J}_x \right\|_2 = \Theta(\sqrt{N_{L-1} N_{L-2}}), \tag{C.173}$$

with probability at least $1 - C\exp(-cN_{L-1}) - 12\exp(-c\log^2 N_{L-1})$ over x and $(W_k)_{k=1}^L$ and x.

Proof. We have that

$$\begin{aligned} \left\| \tilde{J}_x \right\|_2 &= \left\| \tilde{f}_{L-2}(x) \otimes (D_L \tilde{\phi}'(g_{L-1}(x))) - \mathbb{E}_x \left[\tilde{f}_{L-2}(x) \otimes D_L \tilde{\phi}'(g_{L-1}(x)) \right] \right\|_2 \\ &= \left\| \tilde{f}_{L-2}(x) (D_L \tilde{\phi}'(g_{L-1}(x)))^\top - \mathbb{E}_x \left[\tilde{f}_{L-2}(x) (D_L \tilde{\phi}'(g_{L-1}(x)))^\top \right] \right\|_F \\ &= \left\| \tilde{f}_{L-2}(x)^\top (D_L \tilde{\phi}'(g_{L-1}(x))) - A \right\|_F, \end{aligned} \tag{C.174}$$

where A is defined in (C.168). The second equality is justified by the identity $\|u \otimes v\|_2 = \left\| uv^\top \right\|_F$ that holds for any vectors u, v. An application of the triangle inequality gives that

$$\eta - \|A\|_F \le \left\| \tilde{J}_x \right\|_2 \le \eta + \|A\|_F, \quad \text{with } \eta = \left\| \tilde{f}_{L-2}(x) \right\|_2 \left\| D_L \tilde{\phi}'(g_{L-1}(x)) \right\|_2. \tag{C.175}$$

Lemma C.19 gives that

$$\left\| \tilde{f}_{L-2}(x) \right\|_2 = \|f_{L-2}(x) - \mathbb{E}_x [f_{L-2}(x)]\|_2 = \Theta(\sqrt{N_{L-2}}), \tag{C.176}$$

with probability at least $1 - C'\exp(-cN_{L-1})$ over $(W_k)_{k=1}^L$ and x. Furthermore, Lemma C.20 gives that

$$\left\| D_L \tilde{\phi}'(g_{L-1}(x)) \right\|_2 = \|D_L (\phi'(g_{L-1}(x)) - \mathbb{E}_x [\phi'(g_{L-1}(x))])\|_2 = \Theta(\sqrt{N_{L-1}}), \tag{C.177}$$

with probability at least $1 - 10\exp(-c\log^2 N_{L-1}) - C'\exp(-cN_{L-1})$ over x and $(W_k)_{k=1}^L$. By combining (C.175), (C.176), (C.177) and the bound on $\|A\|_F$ provided by Lemma C.23, we conclude that

$$\left\| \tilde{J}_x \right\|_2 = \Theta(\sqrt{N_{L-1} N_{L-2}}), \tag{C.178}$$

with probability at least

$$\begin{aligned} &1 - C'\exp(-cN_{L-1}) - 10\exp(-c\log^2 N_{L-1}) \\ &\quad - C'\exp(-cN_{L-1}) - 2\exp(-c\log^2 N_{L-1}) - C'\exp(-cN_{L-1}) \\ &\ge 1 - C\exp(-cN_{L-1}) - 12\exp(-c\log^2 N_{L-1}), \end{aligned} \tag{C.179}$$

over x and $(W_k)_{k=1}^L$, which gives the desired result. □

Lemma C.25 (Sub-exponential norm of rows of centered Jacobian)**.** *Consider the setting of Theorem 2.6, let $x \sim P_X$, and let $\tilde{J}_x$ be defined as in* (C.172)*. Fix a realization of $(W_k)_{k=1}^L$. Then, with probability at least $1 - 2\exp(-c\log^2 N_{L-1}) - C\exp(-cN_{L-1})$ over this realization (c being a numerical constant), we have that*

$$\left\| \tilde{J}_x \right\|_{\psi_1} = O(\log N_{L-1}). \tag{C.180}$$

Proof. We condition on $\|f_{L-2}(x)\|_{\text{Lip}} = O(1)$ and on $\|\phi'(g_{L-1}(x))\|_{\text{Lip}} = O(1)$. By Lemma C.2, these two conditions hold with probability at least $1 - C\exp(-cN_{L-1})$ over $(W_k)_{k=1}^{L-1}$. Then, we have

$$\begin{aligned}
\left\|\tilde{J}_x\right\|_{\psi_1} &= \sup_{u \text{ s.t. } \|u\|_2=1} \left\|u^\top \tilde{J}_x\right\|_{\psi_1} \\
&= \sup_{U \text{ s.t. } \|U\|_F=1} \left\|\tilde{f}_{L-2}(x)U(D_L\tilde{\phi}'(g_{L-1}(x))) - \mathbb{E}_x\left[\tilde{f}_{L-2}(x)UD_L\tilde{\phi}'(g_{L-1}(x))\right]\right\|_{\psi_1} \\
&\leq C_0 \sup_{U \text{ s.t. } \|U\|_F=1} \|UD_L\|_F \\
&\leq C_0 \|D_L\|_{\text{op}} \\
&\leq C_0 \log N_{L-1},
\end{aligned} \tag{C.181}$$

where the third line follows from Lemma C.5 and the last inequality holds with probability at least $1 - 2\exp(-c\log^2 N_{L-1})$ over W_L by Lemma C.3. Taking into account the initial conditioning, we get the desired result. □

C.4.2 Proof of Proposition 2.8

Proof of Proposition 2.8. Following the notation in Adamczak et al. [2011], we define

$$B := \sup_{z\in\mathbb{R}^n:\|z\|_2=1} \left| \left\| \sum_{i=1}^n z_i \tilde{J}_{i:} \right\|_2^2 - \sum_{i=1}^n z_i^2 \left\|\tilde{J}_{i:}\right\|_2^2 \right|^{\frac{1}{2}}. \tag{C.182}$$

Then, for any $z \in \mathbb{R}^n$ with unit norm, we have that

$$\left\|\tilde{J}z\right\|_2^2 = \left\| \sum_{i=1}^n z_i \tilde{J}_{i:} \right\|_2^2 - \sum_{i=1}^n z_i^2 \left\|\tilde{J}_{i:}\right\|_2^2 + \sum_{i=1}^n z_i^2 \left\|\tilde{J}_{i:}\right\|_2^2 \geq \min_i \left\|\tilde{J}_{i:}\right\|_2^2 - B^2, \tag{C.183}$$

which implies that

$$\lambda_{\min}\left(\tilde{J}\tilde{J}^\top\right) = \inf_{z\in\mathbb{R}^n:\|z\|_2=1} \left\|\tilde{J}z\right\|_2^2 \geq \min_i \left\|\tilde{J}_{i:}\right\|_2^2 - B^2. \tag{C.184}$$

In our case, $\tilde{J}_{i:} \in \mathbb{R}^{N_{L-2}N_{L-1}}$. Notice that this dimension is indicated with n in Theorem 3.2 of Adamczak et al. [2011]. In the statement of the mentioned Theorem, let's fix $r = 1$, $m = n$, and $\theta = (n/(N_{L-1}N_{L-2}))^{1/4} < 1/4$. Then, we have that the condition required to apply Theorem 3.2 is satisfied, *i.e.*

$$n\log^2\left(2\sqrt[4]{\frac{N_{L-2}N_{L-1}}{n}}\right) \leq \sqrt{nN_{L-2}N_{L-1}}, \tag{C.185}$$

where the inequality follows from Assumption 2.5. By combining (C.184) with the upper bound on B given by Theorem 3.2 of Adamczak et al. [2011], the desired result readily follows. □

C.4.3 Proof of Theorem 2.9

Proof of Theorem 2.9. By Lemma C.25, we have that, with probability at least $1 - 2\exp(-c\log^2 N_{L-1}) - C'\exp(-cN_{L-1})$ over $(W_k)_{k=1}^L$, the rows of $\tilde{J}$ are sub-exponential (with respect to the randomness in $(x_i)_{i=1}^n$). In particular, we have that

$$\psi := \max_i \left\|\tilde{J}_{i:}\right\|_{\psi_1} \leq C_1 \log N_{L-1}. \tag{C.186}$$

Furthermore, by Lemma C.24, we have that

$$\left\|\tilde{J}_{i:}\right\|_2 = \Theta(\sqrt{N_{L-2}N_{L-1}}), \tag{C.187}$$

with probability at least $1-p$ over x_i and $(W_k)_{k=1}^L$, where to ease the notation we have defined $p := C'\exp(-c_0 N_{L-1}) + 12\exp(-c_0\log^2 N_{L-1})$. Hence, with probability at least $1-\sqrt{p}$ over $(W_k)_{k=1}^L$, we have that

$$\mathbb{P}_{x_i}\left(c_1\sqrt{N_{L-2}N_{L-1}} \le \left\|\tilde{J}_{i:}\right\|_2 \le c_2\sqrt{N_{L-2}N_{L-1}}\right) \ge 1-\sqrt{p}, \tag{C.188}$$

for some numerical constants $c_2 > c_1 > 0$. In (C.188), we use the symbol $\mathbb{P}_{x_i}$ to highlight that this probability is taken over x_i. For the rest of the argument, we condition on a realization of $(W_k)_{k=1}^L$ s.t. (C.186) and (C.188) hold. Then, by performing a union bound over the samples, we have that

$$\eta_{\mathsf{min}} = \min_i \left\|\tilde{J}_{i:}\right\|_2 \ge c_1\sqrt{N_{L-2}N_{L-1}}, \tag{C.189}$$

and

$$\eta_{\mathsf{max}} = \max_i \left\|\tilde{J}_{i:}\right\|_2 \le c_2\sqrt{N_{L-2}N_{L-1}}, \tag{C.190}$$

with probability at least $1-n\sqrt{p}$ over $(x_i)_{i=1}^n$.

Next, we apply Proposition 2.8 with $K=1$, $K'=c_2$ and

$$\begin{aligned}\Delta &= C_1(\psi K + K')^2 n^{1/4}(N_{L-1}N_{L-2})^{3/4} \\ &\le C_2\log^2 N_{L-1} n^{1/4}(N_{L-1}N_{L-2})^{3/4} \\ &= o(N_{L-1}N_{L-2}).\end{aligned} \tag{C.191}$$

Note that (C.1) in Lemma C.1 gives that $n^{1/4}\cdot\log^2 N_{L-1} = o((N_{L-1}N_{L-2})^{1/4})$, which justifies the last line. Thus, (2.13) implies that

$$\lambda_{\min}\left(\tilde{J}\tilde{J}^\top\right) \ge \eta_{\mathsf{min}}^2 - \Delta \ge c_1 N_{L-2}N_{L-1} - o(N_{L-1}N_{L-2}) = \Theta(N_{L-2}N_{L-1}), \tag{C.192}$$

with probability at least

$$\begin{aligned}1 - \exp\left(-cK\sqrt{n}\log\left(\frac{2N_{L-1}N_{L-2}}{n}\right)\right) - \mathbb{P}\left(\eta_{\mathsf{max}} \ge K'\sqrt{N_{L-1}N_{L-2}}\right) \\ \ge 1 - \exp\left(-c\sqrt{n}\right) - \mathbb{P}\left(\eta_{\mathsf{max}} \ge c_2\sqrt{N_{L-1}N_{L-2}}\right) \\ \ge 1 - \exp\left(-c\sqrt{n}\right) - n\sqrt{p},\end{aligned} \tag{C.193}$$

where the last inequality follows from (C.190). By taking into account the conditioning over $(W_k)_{k=1}^L$ made in order to guarantee (C.186) and (C.188), we conclude that $\lambda_{\min}\left(\tilde{J}\tilde{J}^\top\right) = \Omega(N_{L-1}N_{L-2})$ with probability at least

$$\begin{aligned}&1 - \exp\left(-c\sqrt{n}\right) - n\sqrt{p} - \sqrt{p} - 2\exp(-c\log^2 N_{L-1}) - C'\exp(-cN_{L-1}) \\ &= 1 - \exp\left(-c\sqrt{n}\right) - (n+1)\sqrt{C'\exp(-c_0 N_{L-1}) + 12\exp(-c_0\log^2 N_{L-1})} \\ &\qquad - 2\exp(-c\log^2 N_{L-1}) - C'\exp(-cN_{L-1}) \\ &\ge 1 - \exp\left(-c\sqrt{n}\right) - (n+1)\left(\sqrt{C'}\exp(-c_0 N_{L-1}/2) + \sqrt{12}\exp(-c_0\log^2 N_{L-1}/2)\right) \\ &\qquad - 2\exp(-c\log^2 N_{L-1}) - C'\exp(-cN_{L-1}) \\ &\ge 1 - \exp\left(-c\sqrt{n}\right) - C''n\exp(-c_1 N_{L-1}) - C''n\exp(-c\log^2 N_{L-1}),\end{aligned} \tag{C.194}$$

over $(x_i)_{i=1}^n$ and $(W_k)_{k=1}^L$, which gives the desired result. □

C.5 Proof of the upper bound 2.7

Before giving the proof of the upper bound 2.7, we provide again its statement for the reader's convenience.

Lemma C.26 (Upper bound on the smallest NTK eigenvalue)**.** *Consider the setting of Theorem 2.6, and let K be the NTK Gram matrix* (2.3)*. Then, we have*

$$\lambda_{\min}\left(K\right) = O\left(dN_{L-1}\right), \tag{C.195}$$

with probability at least $1 - C\exp(-cN_{L-1})$ over $(x_i)_{i=1}^n$ and $(W_k)_{k=1}^L$, where c and C are numerical constants.

Proof. By using the expression in (2.8), we have that

$$\lambda_{\min}\left(K\right) = \lambda_{\min}\left(JJ^\top\right) \leq (JJ^\top)_{11} = \sum_{l=0}^{L-1} \|(F_l)_{1:}\|_2^2 \, \|(B_{l+1})_{1:}\|_2^2 . \tag{C.196}$$

An application of Lemma C.16 gives that

$$\|(F_l)_{1:}\|_2^2 = \|f_l(x_1)\|_2^2 = \Theta(N_l), \tag{C.197}$$

with probability at least $1 - C'\exp(-cN_{L-1})$ over $(W_k)_{k=1}^l$ and x_1. We condition on the event such that (C.197) holds for all $l \in \{0, \dots, L-1\}$. This happens with probability at least $1 - C''\exp(-cN_{L-1})$ over $(W_k)_{k=1}^L$ and x_1.

By definition, we have that $\|B_L\|_2 = 1$ and that, for $l \in [L-1]$,

$$\|(B_l)_{1:}\|_2^2 = \left\| \prod_{k=l}^{L-1} \Sigma_k(x_1) W_{k+1} \right\|_2^2 . \tag{C.198}$$

Since $\Sigma_k(x_1) = \mathrm{diag}([\phi'(g_{k,j}(x_1))]_{j=1}^{N_k})$, by Assumption 2.3, we have that

$$\|\Sigma_k(x_1)\|_{\mathsf{op}} \leq M. \tag{C.199}$$

Let us now condition on the following two events: *(i)* $\|W_k\|_{\mathsf{op}} = \Theta(1)$, for all $k \in [L-1]$ (this happens with probability at least $1 - C'\exp(-cN_{L-1})$ over $(W_k)_{k=1}^{L-1}$, see (C.12) in the proof of Lemma C.2), and *(ii)* $\|W_L\|_2 = \Theta(\sqrt{N_{L-1}})$ (this happens with probability at least $1 - \exp(-cN_{L-1})$ over W_L, by Theorem 3.1.1 in Vershynin [2018]). Then, we readily get

$$\|(B_l)_{1:}\|_2^2 = O\left(N_{L-1}\right). \tag{C.200}$$

Taking the intersection of all the events over which we conditioned, we finally obtain

$$\sum_{l=0}^{L-1} \|(F_l)_{1:}\|_2^2 \, \|(B_{l+1})_{1:}\|_2^2 = O\left(N_{L-1} \sum_{l=0}^{L-1} N_l\right) = O\left(dN_{L-1}\right), \tag{C.201}$$

with probability at least $1 - (1 + C' + C'')\exp(-cN_{L-1})$ over $(W_k)_{k=1}^L$ and x_1, where in the last step we have used Assumption 2.4. By combining (C.196) and (C.201), the desired result follows. □

C.6 Proof of Corollary 2.10

Proof of Corollary 2.10. By Theorem 2.6, we have that the smallest eigenvalue of $JJ^\top$ is bounded away from zero with probability at least $1-p$ over $(x_i)_{i=1}^n$ and $(W_k)_{k=1}^L$, where

$$p := C\,ne^{-c\log^2 N_{L-1}} - Ce^{-c\log^2 n}. \tag{C.202}$$

Hence, with probability at least $1-p$ over $(x_i)_{i=1}^n$, there exists a set of parameters θ_0 such that $J(\theta_0)$ has a right inverse. Thus, for all $Y \in \mathbb{R}^n$, there exists θ' such that

$$J(\theta_0)\theta' = \left.\frac{\partial F_L(\theta)}{\partial \theta}\right|_{\theta=\theta_0} \theta' = Y. \tag{C.203}$$

This can also be written, for all $i \in [n]$, as

$$y_i = \left.\frac{\partial f_L(\theta, x_i)}{\partial \theta}\right|_{\theta=\theta_0}^\top \theta' = \lim_{h\to 0} \frac{f_L(\theta_0 + h\theta', x_i) - f_L(\theta_0, x_i)}{h}. \tag{C.204}$$

Then, for all $\varepsilon > 0$, there exists h^* such that, for all $i \in [n]$,

$$|y_i - f^*(x_i)| \le \frac{\varepsilon}{\sqrt{n}}, \tag{C.205}$$

where

$$f^*(x_i) := \frac{f_L(\theta_0 + h^*\theta', x_i) - f_L(\theta_0, x_i)}{h^*}. \tag{C.206}$$

Finally, the desired result follows by noticing that f^* can be implemented by a network with the same depth and twice more neurons at every hidden layer. □

C.7 Proof of Theorem 2.11

Notation for this appendix. In this appendix, we use $J(\theta)$ to denote the Jacobian of the network output F_L, evaluated in θ. We recall that $J(\theta)$ is a matrix with n rows and $\sum_{l=0}^{L-3} N_l N_{l+1} + 2N_{L-2}N_{L-1} + 2N_{L-1}$ columns (for the optimization result, we assume that the $(L-1)$-th layer has an even number of neurons and denote its width as $2N_{L-1}$). Let $K(\theta) = J(\theta)(J(\theta))^\top$ be the associated empirical NTK Gram matrix, and let θ_0 be the initialization defined in (2.18). We also make the dependence on θ explicit for feature vectors and backpropagation terms: the feature vector at the l-th layer with input x_i and network parameter θ is denoted by $f_l(\theta, x_i)$, and the corresponding backpropagation term is denoted by $b_l(\theta, x_i)$, where $b_l(\theta, x_i) = (B_l(\theta))_{i:}$. Finally, we use $W_l(\theta)$ to denote the weights of the l-th layer evaluated at the parameter θ.

A straightforward application of Theorem 2.1 in Oymak and Soltanolkotabi [2019] gives the following proposition.

Proposition C.27. *Consider solving the least-squares optimization problem*

$$\min_\theta \mathcal{L}(\theta) := \frac{1}{2}\min_\theta \|F_L(\theta) - Y\|_2^2, \tag{C.207}$$

by running gradient descent updates of the form $\theta_{t+1} = \theta_t - \eta\nabla\mathcal{L}(\theta_t)$, with some initialization $\tilde{\theta}_0$. Assume there exists $\alpha, \beta \in \mathbb{R}$, such that, if we define $\mathcal{D} = \mathcal{B}(\tilde{\theta}_0, R)$ as the ℓ_2 ball centered in $\tilde{\theta}_0$ with radius R, with

$$R := \frac{4\left\|F_L(\tilde{\theta}_0) - Y\right\|_2}{\alpha}, \tag{C.208}$$

the following holds

$$\forall \theta \in \mathcal{D}: \ \alpha \leq \sigma_{\min}(J(\theta)) \leq \|J(\theta)\|_{\mathsf{op}} \leq \beta, \tag{C.209}$$

$$\forall \theta_1, \theta_2 \in \mathcal{D}: \ \|J(\theta_1) - J(\theta_2)\|_{\mathsf{op}} \leq \frac{\alpha^2}{2\beta}. \tag{C.210}$$

Then, by setting $\eta \leq 1/(2\beta^2)$, we have that, for all $t \geq 1$,

$$\mathcal{L}(\theta_t) \leq \left(1 - \frac{\eta\alpha^2}{2}\right)^t \mathcal{L}(\tilde{\theta}_0). \tag{C.211}$$

In order to apply this proposition with initialization $\tilde{\theta}_0 = \theta_0$, we need to prove that the necessary assumptions hold. We will do so by showing the following intermediate results:

- Lemma C.28 shows that, at the initial point θ_0, the network output is 0 and the smallest NTK eigenvalue is lower bounded.
- Lemma C.29 gives a tight estimate on the operator norm of the weights inside a ball $\mathcal{D}$ centered at θ_0 and with radius $R = o(1)$.
- Lemma C.30 gives an upper bound on the distance between a feature vector in $\mathcal{D}$ and the feature vector at θ_0.
- Lemma C.31 gives upper bounds on the ℓ_2 norm and the ℓ_2 distance between feature vectors in $\mathcal{D}$.
- Lemmas C.32 and C.33 give upper bounds on the ℓ_2 norm and the ℓ_2 distance of backpropagation terms in $\mathcal{D}$, respectively.
- Lemma C.34 gives an upper bound on the difference in operator norm between Jacobians in $\mathcal{D}$.
- Finally, Lemma C.35 gives upper and lower bounds on the NTK spectrum in $\mathcal{D}$.

Lemma C.28 (Network output and smallest NTK eigenvalue at initialization)**.** *Let θ_0 be defined in* (2.18)*. Then, we have that, for all $x \in \mathbb{R}^d$,*

$$f_L(x, \theta_0) = 0. \tag{C.212}$$

Furthermore, we have that

$$\sigma_{\min}(J(\theta_0)) \geq c_1\sqrt{\gamma N_{L-2}N_{L-1}}), \tag{C.213}$$

with probability at least $1 - C\,ne^{-c\log^2 N_{L-1}} - Ce^{-c\log^2 n}$ over $(x_i)_{i=1}^n \sim_{\text{i.i.d.}} P_X$ and θ_0, where c, c_1 and C are numerical constants.

Proof. By definition (2.18) of the initialization θ_0, we have that

$$\begin{aligned} f_L(x, \theta_0) &= (W_L^{(1)}(\theta_0))^\top \phi((W_{L-1}^{(1)}(\theta_0))^\top f_{L-2}(\theta_0, x)) \\ &\quad + (W_L^{(2)}(\theta_0))^\top \phi((W_{L-1}^{(2)}(\theta_0))^\top f_{L-2}(\theta_0, x)) \\ &= (W_L^{(1)}(\theta_0))^\top \phi((W_{L-1}^{(1)}(\theta_0))^\top f_{L-2}(\theta_0, x)) \\ &\quad + (-W_L^{(1)}(\theta_0))^\top \phi((W_{L-1}^{(1)}(\theta_0))^\top f_{L-2}(\theta_0, x)) \\ &= 0, \end{aligned} \tag{C.214}$$

where in the second equality we use that $W_{L-1}^{(2)}(\theta_0) = W_{L-1}^{(1)}(\theta_0)$ and that $W_L^{(2)}(\theta_0) = -W_L^{(1)}(\theta_0)$. This proves (C.212).

Let us now compute the Jacobian at initialization $J(\theta_0)$. For $l \in [L-2]$, we have that

$$\begin{aligned}
\frac{\partial f_L(x)}{\partial (W_l)_{ij}}\bigg|_{\theta=\theta_0} &= (W_L^{(1)}(\theta_0))^\top \Bigg(\phi'\left((W_{L-1}^{(1)}(\theta_0))^\top f_{L-2}(\theta_0, x)\right) \left((W_{L-1}^{(1)}(\theta_0))^\top \frac{\partial f_{L-2}(\theta, x)}{\partial (W_l)_{ij}}\bigg|_{\theta=\theta_0} \right) \Bigg) \\
&+ (W_L^{(2)}(\theta_0))^\top \left(\phi'\left((W_{L-1}^{(2)}(\theta_0))^\top f_{L-2}(\theta_0, x)\right) \left((W_{L-1}^{(2)}(\theta_0))^\top \frac{\partial f_{L-2}(\theta, x)}{\partial (W_l)_{ij}}\bigg|_{\theta=\theta_0} \right) \right) \\
&= (W_L^{(1)}(\theta_0))^\top \Bigg(\phi'\left((W_{L-1}^{(1)}(\theta_0))^\top f_{L-2}(\theta_0, x)\right) \left((W_{L-1}^{(1)}(\theta_0))^\top \frac{\partial f_{L-2}(\theta, x)}{\partial (W_l)_{ij}}\bigg|_{\theta=\theta_0} \right) \Bigg) \\
&- (W_L^{(1)}(\theta_0))^\top \left(\phi'\left((W_{L-1}^{(1)}(\theta_0))^\top f_{L-2}(\theta_0, x)\right) \left((W_{L-1}^{(1)}(\theta_0))^\top \frac{\partial f_{L-2}(\theta, x)}{\partial (W_l)_{ij}}\bigg|_{\theta=\theta_0} \right) \right) \\
&= 0,
\end{aligned} \tag{C.215}$$

where in the second equality we use again that $W_{L-1}^{(2)}(\theta_0) = W_{L-1}^{(1)}(\theta_0)$ and that $W_L^{(2)}(\theta_0) = -W_L^{(1)}(\theta_0)$.

Let us define $f_{L-1}^{(k)}(\theta, x) := \phi((W_{L-1}^{(k)}(\theta))^\top f_{L-2}(\theta, x))$ for $k \in \{1, 2\}$. Then, for the $(L-1)$-th layer, by isolating the computation over $W_{L-1}^{(1)}$, we have that

$$\begin{aligned}
\frac{\partial f_L(\theta, x)}{\partial (W_{L-1}^{(1)})_{ij}}\bigg|_{\theta=\theta_0} &= (W_L^{(1)}(\theta_0))^\top \frac{\partial f_{L-1}^{(1)}(\theta, x)}{\partial (W_{L-1}^{(1)})_{ij}}\bigg|_{\theta=\theta_0} + (W_L^{(2)}(\theta_0))^\top \frac{\partial f_{L-1}^{(2)}(\theta, x)}{\partial (W_{L-1}^{(1)})_{ij}}\bigg|_{\theta=\theta_0} \\
&= (W_L^{(1)}(\theta_0))^\top \frac{\partial f_{L-1}^{(1)}(\theta, x)}{\partial (W_{L-1}^{(1)})_{ij}}\bigg|_{\theta=\theta_0} \\
&=: J^{(1)}(\theta_0),
\end{aligned} \tag{C.216}$$

where we use that $f_{L-1}^{(2)}(\theta, x)$ does not depend on the parameters $W_{L-1}^{(1)}$. Proceeding in the same way and using that $W_{L-1}^{(2)}(\theta_0) = W_{L-1}^{(1)}(\theta_0)$ and $W_L^{(2)}(\theta_0) = -W_L^{(1)}(\theta_0)$, we also obtain that

$$\frac{\partial f_L(\theta, x)}{\partial (W_{L-1}^{(2)})_{ij}}\bigg|_{\theta=\theta_0} = -J^{(1)}(\theta_0). \tag{C.217}$$

Finally, by observing that $f_L(\theta, x) = (W_L^{(1)})^\top f_{L-1}^{(1)}(\theta, x) + (W_L^{(2)})^\top f_{L-1}^{(2)}(\theta, x)$, we deduce

$$\frac{\partial f_L(\theta, x)}{\partial (W_L^{(k)})_i}\bigg|_{\theta=\theta_0} = \left(f_{L-1}^{(k)}(\theta_0, x) \right)_i, \quad \text{for } k \in \{1, 2\}. \tag{C.218}$$

Hence, the NTK at initialization $K(\theta_0)$ can be expressed as

$$\begin{aligned}
K(\theta_0) = J^{(1)}(\theta_0)(J^{(1)}(\theta_0))^\top + J^{(1)}(\theta_0)(J^{(1)}(\theta_0))^\top \\
+ F_{L-1}^{(1)}(\theta_0)(F_{L-1}^{(1)}(\theta_0))^\top + F_{L-1}^{(2)}(\theta_0)(F_{L-1}^{(2)}(\theta_0))^\top.
\end{aligned} \tag{C.219}$$

Note that, by construction, $J^{(1)}(\theta_0)$ has the same distribution of J_{L-2}, whose rows are given by (2.10). Therefore, by combining the results from Theorem 2.7 and Theorem 2.9, we conclude that

$$\sigma_{\min}(J(\theta_0)) \geq 2\sqrt{\gamma}\sigma_{\min}(J_{L-2}) \geq c_1\sqrt{\gamma N_{L-2}N_{L-1}}, \tag{C.220}$$

with probability at least $1 - C\,ne^{-c\log^2 N_{L-1}} - Ce^{-c\log^2 n}$ over $(x_i)_{i=1}^n$ and θ_0, where c_1, C are numerical constants. This proves (C.213), and concludes the proof of the lemma. □

Lemma C.29 (Operator norm of weights in D)**.** *Let θ_0 be defined in* (2.18)*, let $\mathcal{D} = \mathcal{B}(\theta_0, R)$ and assume that $R = o(1)$. Then, for any $l \in [L-1]$,*

$$\sup_{\theta\in\mathcal{D}} \|W_l(\theta)\|_{\mathsf{op}} = O(1), \tag{C.221}$$

with probability at least $1 - 2\exp(-cN_{L-1})$ over $W_l(\theta_0)$.

Proof. By Weyl's theorem, we have that, for all $l \in [L-2]$,

$$\begin{aligned}
\sup_{\theta\in\mathcal{D}} \|W_l(\theta)\|_{\mathsf{op}} &\le \|W_l(\theta_0)\|_{\mathsf{op}} + \sup_{\theta\in\mathcal{D}} \|W_l(\theta) - W_l(\theta_0)\|_{\mathsf{op}} \\
&\le \|W_l(\theta_0)\|_{\mathsf{op}} + \sup_{\theta\in\mathcal{D}} \|W_l(\theta) - W_l(\theta_0)\|_F \\
&\le \|W_l(\theta_0)\|_{\mathsf{op}} + \sup_{\theta\in\mathcal{D}} \|\theta - \theta_0\|_2 \\
&= \|W_l(\theta_0)\|_{\mathsf{op}} + o(1) \\
&= O(1),
\end{aligned} \tag{C.222}$$

where in the fourth line we use that $R = o(1)$, and the result of the last line holds with probability at least $1 - 2\exp(-cN_{L-1})$ over $W_l(\theta_0)$ by Theorem 4.4.5 of Vershynin [2018]. By following the same argument, we have that, with probability at least $1 - 2\exp(-cN_{L-1})$ over $W_{L-1}(\theta_0)$,

$$\sup_{\theta\in\mathcal{D}} \left\|W_{L-1}^{(k)}(\theta)\right\|_{\mathsf{op}} = O(1), \quad \text{for } k \in \{1,2\}, \tag{C.223}$$

which readily implies that $\sup_{\theta\in\mathcal{D}} \|W_{L-1}(\theta)\|_{\mathsf{op}} = O(1)$ and concludes the proof. □

Lemma C.30 (Distance of features in D from initialization)**.** *Let θ_0 be defined in* (2.18)*, $x \sim P_X$, $\mathcal{D} = \mathcal{B}(\theta_0, R)$ and assume that $R = o(1)$. Then, for any $0 \le l \le L-1$, we have*

$$\sup_{\theta\in\mathcal{D}} \|f_l(\theta, x) - f_l(\theta_0, x)\|_2 \le C\,R\sqrt{d}, \tag{C.224}$$

with probability at least $1 - C\exp(-cN_{L-1})$ over $(W_k(\theta_0))_{k=1}^l$ and x, where c, C are numerical constants.

Proof. We prove the claim by induction over l. For the base case, we have $f_0(\theta, x) = f_0(\theta_0, x)$, hence (C.224) holds with probability 1.

For the induction case, let $l > 0$. Then,

$$\begin{aligned}
\sup_{\theta\in\mathcal{D}} \|f_l(\theta, x) - f_l(\theta_0, x)\|_2 &= \sup_{\theta\in\mathcal{D}} \left\|\phi\left((W_l(\theta))^\top f_{l-1}(\theta, x)\right) - \phi\left((W_l(\theta_0))^\top f_{l-1}(\theta_0, x)\right)\right\|_2 \\
&\le M \sup_{\theta\in\mathcal{D}} \left\|(W_l(\theta))^\top f_{l-1}(\theta, x) - (W_l(\theta_0))^\top f_{l-1}(\theta_0, x)\right\|_2 \\
&\le M \sup_{\theta\in\mathcal{D}} \left\|(W_l(\theta))^\top f_{l-1}(\theta, x) - (W_l(\theta))^\top f_{l-1}(\theta_0, x)\right\|_2 \\
&\quad + M \sup_{\theta\in\mathcal{D}} \left\|(W_l(\theta))^\top f_{l-1}(\theta_0, x) - (W_l(\theta_0))^\top f_{l-1}(\theta_0, x)\right\|_2 \\
&\le M \sup_{\theta\in\mathcal{D}} \|W_l(\theta)\|_{\mathsf{op}} \sup_{\theta\in\mathcal{D}} \|f_{l-1}(\theta, x) - f_{l-1}(\theta_0, x)\|_2 \\
&\quad + M \sup_{\theta\in\mathcal{D}} \|W_l(\theta) - W_l(\theta_0)\|_{\mathsf{op}} \|f_{l-1}(\theta_0, x)\|_2 .
\end{aligned} \tag{C.225}$$

By Lemma C.29, we have that

$$\sup_{\theta\in\mathcal{D}} \|W_l(\theta)\|_{\mathsf{op}} = O\left(1\right), \tag{C.226}$$

with probability at least $1-2\exp(-cN_{L-1})$ over $W_l(\theta_0)$. By inductive hypothesis, we have

$$\sup_{\theta\in\mathcal{D}} \|f_{l-1}(\theta,x) - f_{l-1}(\theta_0,x)\|_2 \leq C\,R\sqrt{d}, \tag{C.227}$$

with probability at least $1-C\exp(-cN_{L-1})$ over $(W_k(\theta_0))_{k=1}^{l-1}$ and x. Clearly, we also have that

$$\sup_{\theta\in\mathcal{D}} \|W_l(\theta) - W_l(\theta_0)\|_{\mathsf{op}} \leq \sup_{\theta\in\mathcal{D}} \|W_l(\theta) - W_l(\theta_0)\|_F \leq \sup_{\theta\in\mathcal{D}} \|\theta-\theta_0\| \leq R. \tag{C.228}$$

Furthermore, an application of Lemma C.16 gives that

$$\|f_{l-1}(\theta_0,x)\|_2 = \Theta(\sqrt{N_{l-1}}) = O\left(\sqrt{d}\right), \tag{C.229}$$

with probability at least $1-C\exp(-cN_{L-1})$ over $(W_k(\theta_0))_{k=1}^{l-1}$ and x. By combining (C.225), (C.226), (C.227), (C.228) and (C.229), we obtain that

$$\sup_{\theta\in\mathcal{D}} \|f_l(\theta,x) - f_l(\theta_0,x)\|_2 \leq C\,R\sqrt{d}, \tag{C.230}$$

with probability at least $1-C\exp(-cN_{L-1})$ over $(W_k(\theta_0))_{k=1}^{l}$ and x, which completes the proof. □

Lemma C.31 (L2 norm and distance of features in D)**.** *Let θ_0 be defined in* (2.18)*, $x\sim P_X$, $\mathcal{D}=\mathcal{B}(\theta_0,R)$ and assume that $R=o(1)$. Then, for any $0\leq l\leq L-1$, we have*

$$\sup_{\theta\in\mathcal{D}} \|f_l(\theta,x)\|_2 = O\left(\sqrt{d}\right), \tag{C.231}$$

with probability at least $1-C\exp(-cN_{L-1})$ over $(W_k(\theta_0))_{k=1}^{l}$ and x, where c,C are numerical constants. Furthermore,

$$\sup_{\theta_1,\theta_2\in\mathcal{D}} \|f_l(\theta_1,x) - f_l(\theta_2,x)\|_2 \leq C\,R\sqrt{d}, \tag{C.232}$$

with probability at least $1-C\exp(-cN_{L-1})$ over $(W_k(\theta_0))_{k=1}^{l}$ and x.

Proof. The first statement follows from the chain of inequalities below:

$$\sup_{\theta\in\mathcal{D}} \|f_l(\theta,x)\|_2 \leq \|f_l(\theta_0,x)\|_2 + \sup_{\theta\in\mathcal{D}} \|f_l(\theta) - f_l(\theta_0)\|_2 \leq C\sqrt{N_l} + C\,R\sqrt{d}, \tag{C.233}$$

where the second inequality holds with probability at least $1-C\exp(-cN_{L-1})$ over $(W_k(\theta_0))_{k=1}^{l}$ and x by combining Lemma C.16 and Lemma C.30.

We prove the second statement by induction over l. For the base case, we have $f_0(\theta_1,x) = f_0(\theta_2,x)$, hence (C.232) holds with probability 1.

For the induction case, let $l > 0$. Then,

$$\begin{aligned}
\sup_{\theta_1,\theta_2\in\mathcal{D}} \|f_l(\theta_1,x)-f_l(\theta_2,x)\|_2 &= \sup_{\theta_1,\theta_2\in\mathcal{D}} \left\|\phi\left((W_l(\theta_1))^\top f_{l-1}(\theta_1,x)\right)-\phi\left((W_l(\theta_2))^\top f_{l-1}(\theta_2,x)\right)\right\|_2 \\
&\leq M \sup_{\theta_1,\theta_2\in\mathcal{D}} \left\|(W_l(\theta_1))^\top f_{l-1}(\theta_1,x) - (W_l(\theta_2))^\top f_{l-1}(\theta_2,x)\right\|_2 \\
&\leq M \sup_{\theta_1,\theta_2\in\mathcal{D}} \left\|(W_l(\theta_1))^\top f_{l-1}(\theta_1,x) - (W_l(\theta_1))^\top f_{l-1}(\theta_2,x)\right\|_2 \\
&\quad + M \sup_{\theta_1,\theta_2\in\mathcal{D}} \left\|(W_l(\theta_1))^\top f_{l-1}(\theta_2,x) - (W_l(\theta_2))^\top f_{l-1}(\theta_2,x)\right\|_2 \\
&\leq M \sup_{\theta_1\in\mathcal{D}} \|W_l(\theta_1)\|_{\mathsf{op}} \sup_{\theta_1,\theta_2\in\mathcal{D}} \|f_{l-1}(\theta_1,x) - f_{l-1}(\theta_2,x)\|_2 \\
&\quad + M \sup_{\theta_1,\theta_2\in\mathcal{D}} \|W_l(\theta_1) - W_l(\theta_2)\|_{\mathsf{op}} \sup_{\theta_2\in\mathcal{D}} \|f_{l-1}(\theta_2,x)\|_2 .
\end{aligned} \tag{C.234}$$

By Lemma C.29, we have that

$$\sup_{\theta_1\in\mathcal{D}} \|W_l(\theta_1)\|_{\mathsf{op}} = O(1), \tag{C.235}$$

with probability at least $1 - 2\exp(-cN_{L-1})$ over $W_l(\theta_0)$. By inductive hypothesis, we have

$$\sup_{\theta_1,\theta_2\in\mathcal{D}} \|f_{l-1}(\theta_1,x) - f_{l-1}(\theta_2,x)\|_2 \leq C\,R\sqrt{d}, \tag{C.236}$$

with probability at least $1 - C\exp(-cN_{L-1})$ over $(W_k(\theta_0))_{k=1}^{l-1}$ and x. Clearly, we also have that

$$\sup_{\theta_1,\theta_2\in\mathcal{D}} \|W_l(\theta_1) - W_l(\theta_2)\|_{\mathsf{op}} \leq \sup_{\theta_1,\theta_2\in\mathcal{D}} \|W_l(\theta_1) - W_l(\theta_2)\|_F \leq \sup_{\theta_1,\theta_2\in\mathcal{D}} \|\theta_1 - \theta_2\| \leq R. \tag{C.237}$$

Furthermore, by using (C.231), we have that

$$\sup_{\theta_2\in\mathcal{D}} \|f_{l-1}(\theta_2,x)\|_2 \leq C\,R\sqrt{d}, \tag{C.238}$$

with probability at least $1 - C\exp(-cN_{L-1})$ over $(W_k(\theta_0))_{k=1}^{l-1}$ and x. By combining (C.234), (C.235), (C.236), (C.237) and (C.238), we obtain that

$$\sup_{\theta_1,\theta_2\in\mathcal{D}} \|f_l(\theta_1,x) - f_l(\theta_2,x)\|_2 \leq C\,R\sqrt{d}, \tag{C.239}$$

with probability at least $1 - C\exp(-cN_{L-1})$ over $(W_k(\theta_0))_{k=1}^{l}$ and x, which completes the proof. □

Lemma C.32 (L2 norm of backpropagation in D)**.** *Let θ_0 be defined in* (2.18)*, $x \sim P_X$, and $\mathcal{D} = \mathcal{B}(\theta_0, R)$. Assume that $R = o(1)$ and that $\gamma > 1$. Then, for any $l \in [L]$, we have*

$$\sup_{\theta\in\mathcal{D}} \|b_l(\theta,x)\|_2 \leq C\sqrt{\gamma \cdot N_{L-1}}, \tag{C.240}$$

with probability at least $1 - C\exp(-cN_{L-1})$ over $(W_k(\theta_0))_{k=l+1}^{L}$, where c, C are numerical constants.

Proof. We prove the claim by induction on $l \in \{L, L-1, \dots, 1\}$. For the base case, we have that $\|b_L(\theta, x)\|_2 = 1$, hence (C.240) clearly holds.

For the induction case, pick $l \in [L-1]$. Then,

$$\begin{aligned}
\sup_{\theta\in\mathcal{D}} \|b_l(\theta,x)\|_2 &= \sup_{\theta\in\mathcal{D}} \left\| \left(\prod_{k=l}^{L-2} \Sigma_k(\theta,x) W_{k+1}(\theta) \right) \Sigma_{L-1}(\theta,x) W_L(\theta) \right\|_2 \\
&\le \sup_{\theta\in\mathcal{D}} \left\| \left(\prod_{k=l}^{L-2} \Sigma_k(\theta,x) W_{k+1}(\theta) \right) \Sigma_{L-1}(\theta,x) \right\|_{\mathsf{op}} \sup_{\theta\in\mathcal{D}} \|W_L(\theta)\|_2 \\
&\le \left(\prod_{k=l}^{L-2} \sup_{\theta\in\mathcal{D}} \|\Sigma_k(\theta,x)\|_{\mathsf{op}} \sup_{\theta\in\mathcal{D}} \|W_{k+1}(\theta)\|_{\mathsf{op}} \right) \sup_{\theta\in\mathcal{D}} \|\Sigma_{L-1}(\theta,x)\|_{\mathsf{op}} \sup_{\theta\in\mathcal{D}} \|W_L(\theta)\|_2 \\
&\le M^{L-l} \left(\prod_{k=l+1}^{L-1} \sup_{\theta\in\mathcal{D}} \|W_k(\theta)\|_{\mathsf{op}} \right) \left(\|W_L(\theta_0)\|_2 + \sup_{\theta\in\mathcal{D}} \|W_L(\theta) - W_L(\theta_0)\|_2 \right) \\
&\le C\, M^{L-l} (\|W_L(\theta_0)\|_2 + \sup_{\theta\in\mathcal{D}} \|\theta - \theta_0\|_2) \\
&\le C\, M^{L-l} (\sqrt{\gamma N_{L-1}} + \sup_{\theta\in\mathcal{D}} \|\theta - \theta_0\|_2) \\
&= C\sqrt{\gamma N_{L-1}}.
\end{aligned} \tag{C.241}$$

Here, the fourth line follows from Assumption 2.3, which gives $\sup_{\theta\in\mathcal{D}} \|\Sigma_k(\theta,x)\|_{\mathsf{op}} \le M$; the fifth line holds with probability $1 - C\exp(-cN_{L-1})$ over $(W_k(\theta_0))_{k=l+1}^{L-1}$ by Lemma C.29; the sixth line holds with probability at least $1 - \exp(-cN_{L-1})$ over $W_L(\theta_0)$ by Theorem 3.1.1 in Vershynin [2018]; and the last line follows from $R = o(1)$. Taking the intersection of these events gives the desired result. □

Lemma C.33 (L2 distance of backpropagation in D)**.** *Let θ_0 be defined in* (2.18)*, $x \sim P_X$, and $\mathcal{D} = \mathcal{B}(\theta_0, R)$. Assume that $R = o(1)$ and that $\gamma > 1$. Then, for any $l \in [L]$, we have*

$$\sup_{\theta_1,\theta_2\in\mathcal{D}} \|b_l(\theta_1,x) - b_l(\theta_2,x)\|_2 \le C\, R \sqrt{\gamma d N_{L-1}}, \tag{C.242}$$

with probability at least $1 - C\exp(-cN_{L-1})$ over $(W_k(\theta_0))_{k=l+1}^{L}$ and x, where c, C are numerical constants.

Proof. We prove the claim by induction on $l \in \{L, L-1, \dots, 1\}$. For the base case, $b_L(\theta, x)$ does not depend on θ, hence (C.242) clearly holds. For the induction case, pick $l \in [L-1]$.

Then,

$$
\begin{aligned}
&\sup_{\theta_1,\theta_2\in\mathcal{D}} \|b_l(\theta_1,x)-b_l(\theta_2,x)\|_2 \\
&\quad= \sup_{\theta_1,\theta_2\in\mathcal{D}} \|\Sigma_l(\theta_1,x)W_{l+1}(\theta_1)b_{l+1}(\theta_1,x)-\Sigma_l(\theta_2,x)W_{l+1}(\theta_2)b_{l+1}(\theta_2,x)\|_2 \\
&\quad\le \sup_{\theta_1,\theta_2\in\mathcal{D}} \|\Sigma_l(\theta_1,x)W_{l+1}(\theta_1)b_{l+1}(\theta_1,x)-\Sigma_l(\theta_1,x)W_{l+1}(\theta_1)b_{l+1}(\theta_2,x)\|_2 \\
&\qquad+ \sup_{\theta_1,\theta_2\in\mathcal{D}} \|\Sigma_l(\theta_1,x)W_{l+1}(\theta_1)b_{l+1}(\theta_2,x)-\Sigma_l(\theta_2,x)W_{l+1}(\theta_2)b_{l+1}(\theta_2,x)\|_2 \\
&\quad\le \sup_{\theta_1\in\mathcal{D}} \|\Sigma_l(\theta_1,x)\|_{\mathsf{op}} \sup_{\theta_1\in\mathcal{D}} \|W_{l+1}(\theta_1)\|_{\mathsf{op}} \sup_{\theta_1,\theta_2\in\mathcal{D}} \|b_{l+1}(\theta_1,x)-b_{l+1}(\theta_2,x)\|_2 \\
&\qquad+ \sup_{\theta_1,\theta_2\in\mathcal{D}} \|\Sigma_l(\theta_1,x)W_{l+1}(\theta_1)-\Sigma_l(\theta_2,x)W_{l+1}(\theta_2)\|_{\mathsf{op}} \sup_{\theta_2\in\mathcal{D}} \|b_{l+1}(\theta_2,x)\|_2 \\
&\quad\le \sup_{\theta_1\in\mathcal{D}} \|\Sigma_l(\theta_1,x)\|_{\mathsf{op}} \sup_{\theta_1\in\mathcal{D}} \|W_{l+1}(\theta_1)\|_{\mathsf{op}} \sup_{\theta_1,\theta_2\in\mathcal{D}} \|b_{l+1}(\theta_1,x)-b_{l+1}(\theta_2,x)\|_2 \\
&\qquad+ \sup_{\theta_1,\theta_2\in\mathcal{D}} \|\left(\Sigma_l(\theta_1,x)-\Sigma_l(\theta_2,x)\right)W_{l+1}(\theta_2)\|_{\mathsf{op}} \sup_{\theta_2\in\mathcal{D}} \|b_{l+1}(\theta_2,x)\|_2 \\
&\qquad+ \sup_{\theta_1,\theta_2\in\mathcal{D}} \|\Sigma_l(\theta_2,x)\left(W_{l+1}(\theta_1)-W_{l+1}(\theta_2)\right)\|_{\mathsf{op}} \sup_{\theta_2\in\mathcal{D}} \|b_{l+1}(\theta_2,x)\|_2 .
\end{aligned}
\tag{C.243}
$$

Furthermore, we have that the following results hold.

(i) By Assumption 2.3 and Lemma C.29,

$$\sup_{\theta_1\in\mathcal{D}} \|\Sigma_l(\theta_1,x)\|_{\mathsf{op}} \sup_{\theta_1\in\mathcal{D}} \|W_{l+1}(\theta_1)\|_{\mathsf{op}} = O(1),$$

with probability $1-2\exp(-cN_{L-1})$ over $W_{l+1}(\theta_0)$;

(ii) By inductive hypothesis,

$$\sup_{\theta_1,\theta_2\in\mathcal{D}} \|b_{l+1}(\theta_1,x)-b_{l+1}(\theta_2,x)\|_2 \le C\,R\sqrt{\gamma d N_{L-1}},$$

with probability at least $1-C\exp(-cN_{L-1})$ over $(W_k(\theta_0))_{k=l+2}^{L}$ and x;

(iii) By the same argument of the second statement in Lemma C.31 and again Lemma C.29,

$$
\begin{aligned}
&\sup_{\theta_1,\theta_2\in\mathcal{D}} \|\left(\Sigma_l(\theta_1,x)-\Sigma_l(\theta_2,x)\right)W_{l+1}(\theta_2)\|_{\mathsf{op}} \\
&\qquad\le \sup_{\theta_1,\theta_2\in\mathcal{D}} \|\phi'\left(g_l(\theta_1,x)\right)-\phi'\left(g_l(\theta_2,x)\right)\|_2 \sup_{\theta_2\in\mathcal{D}} \|W_{l+1}(\theta_2)\|_{\mathsf{op}} \\
&\qquad\le C\,R\sqrt{d},
\end{aligned}
$$

with probability at least $1-C\exp(-cN_{L-1})$ over $(W_k(\theta_0))_{k=1}^{l+1}$ and x;

(iv) By Lemma C.32,

$$\sup_{\theta_2\in\mathcal{D}} \|b_{l+1}(\theta_2,x)\|_2 \le C\sqrt{\gamma N_{L-1}},$$

with probability at least $1-C\exp(-cN_{L-1})$ over $(W_k(\theta_0))_{k=l+1}^{L}$;

(v) By Assumption 2.3,

$$\sup_{\theta_1,\theta_2\in\mathcal{D}} \|\Sigma_l(\theta_2,x)\left(W_{l+1}(\theta_1)-W_{l+1}(\theta_2)\right)\|_{\mathsf{op}} \le C\,R.$$

By combining (i)-(v) with (C.243), we conclude that

$$\sup_{\theta_1,\theta_2\in\mathcal{D}} \|b_l(\theta_1,x)-b_l(\theta_2,x)\|_2 \leq C\,R\sqrt{\gamma d N_{L-1}}, \tag{C.244}$$

with probability at least $1-C\exp(-cN_{L-1})$ over x and $(W_k(\theta_0))_{k=1}^L$, which concludes the proof. □

Lemma C.34 (Difference of Jacobians in D). *Let θ_0 be defined in* (2.18)*, $x\sim P_X$, and $\mathcal{D}=\mathcal{B}(\theta_0,R)$. Assume that $R=o(1)$ and that $\gamma>1$. Then, we have*

$$\sup_{\theta_1,\theta_2\in\mathcal{D}} \|J(\theta_1)-J(\theta_2)\|_{\mathsf{op}} \leq C\,R\,d\sqrt{\gamma N_{L-1}n}, \tag{C.245}$$

with probability at least $1-Cn\exp(-cN_{L-1})$ over $(x_i)_{i=1}^n$ and θ_0, where c,C are numerical constants.

Proof. Pick $i\in[n]$. Then, we have

$$\begin{aligned}
&\sup_{\theta_1,\theta_2\in\mathcal{D}} \|(J(\theta_1))_{i:}-(J(\theta_2))_{i:}\|_2^2 \\
&\quad\leq \sum_{l=0}^{L-1} \sup_{\theta_1,\theta_2\in\mathcal{D}} \|(F_l(\theta_1))_{:i}\otimes(B_{l+1}(\theta_1))_{:i}-(F_l(\theta_2))_{:i}\otimes(B_{l+1}(\theta_2))_{:i}\|_2^2 \\
&\quad= \sum_{l=0}^{L-1} \sup_{\theta_1,\theta_2\in\mathcal{D}} \|f_l(\theta_1,x_i)\otimes b_{l+1}(\theta_1,x_i)-f_l(\theta_2,x_i)\otimes b_{l+1}(\theta_2,x_i)\|_2^2 \\
&\quad\leq \sum_{l=0}^{L-1} \sup_{\theta_1,\theta_2\in\mathcal{D}} \|(f_l(\theta_1,x_i)-f_l(\theta_2,x_i))\otimes b_{l+1}(\theta_1,x_i)\|_2^2 \\
&\qquad+\sum_{l=0}^{L-1} \sup_{\theta_1,\theta_2\in\mathcal{D}} \|f_l(\theta_2,x_i)\otimes(b_{l+1}(\theta_1,x_i)-b_{l+1}(\theta_1,x_i))\|_2^2 \\
&\quad\leq \sum_{l=0}^{L-1} \sup_{\theta_1,\theta_2\in\mathcal{D}} \|f_l(\theta_1,x_i)-f_l(\theta_2,x_i)\|_2^2 \sup_{\theta_1\in\mathcal{D}} \|b_{l+1}(\theta_1,x_i)\|_2^2 \\
&\qquad+\sum_{l=0}^{L-1} \sup_{\theta_2\in\mathcal{D}} \|f_l(\theta_2,x_i)\|_2^2 \sup_{\theta_1,\theta_2\in\mathcal{D}} \|b_{l+1}(\theta_1,x_i)-b_{l+1}(\theta_1,x_i)\|_2^2 .
\end{aligned} \tag{C.246}$$

Since $x_i\sim P_X$, we can merge together the results from Lemmas C.30, C.31, C.32 and C.33 and obtain

$$\sup_{\theta_1,\theta_2\in\mathcal{D}} \|(J(\theta_1))_{i:}-(J(\theta_2))_{i:}\|_2^2 \leq C\gamma R^2 d^2 N_{L-1}, \tag{C.247}$$

with probability at least $1-C\exp(-cN_{L-1})$ over x_i and θ_0.

Therefore, we have

$$\begin{aligned}
\sup_{\theta_1,\theta_2\in\mathcal{D}} \|J(\theta_1)-J(\theta_2)\|_{\mathsf{op}} &\leq \sup_{\theta_1,\theta_2\in\mathcal{D}} \|J(\theta_1)-J(\theta_2)\|_F \\
&\leq \sqrt{\sum_{i=1}^n \sup_{\theta_1,\theta_2\in\mathcal{D}} \|(J(\theta_1))_{i:}-(J(\theta_2))_{i:}\|_2^2} \\
&\leq C\,R\,d\sqrt{\gamma N_{L-1}n},
\end{aligned} \tag{C.248}$$

with probability $1-Cn\exp(-cN_{L-1})$ over $(x_i)_{i=1}^n$ and θ_0. □

Lemma C.35 (NTK spectrum in D). *Let θ_0 be defined in* (2.18), *$x \sim P_X$, and $\mathcal{D} = \mathcal{B}(\theta_0, R)$. Assume that $R = o(1)$ and that $\gamma > 1$. Then, we have*

$$\sup_{\theta \in \mathcal{D}} \|K(\theta)\|_{\mathsf{op}} \leq C\,\gamma\,n\,d\,N_{L-1}, \tag{C.249}$$

with probability at least $1 - C\,n\exp(-cN_{L-1})$ over θ_0 and $(x_i)_{i=1}^n$, where c, C are numerical constants. Furthermore,

$$\inf_{\theta \in \mathcal{D}} \sigma_{\mathsf{min}}(J(\theta)) \geq c_1\sqrt{\gamma N_{L-2}N_{L-1}} - C_1\,R\,d\sqrt{\gamma N_{L-1}n}, \tag{C.250}$$

with probability at least $1 - C\,ne^{-c\log^2 N_{L-1}} - Ce^{-c\log^2 n}$ over θ_0 and $(x_i)_{i=1}^n$, where c_1, C_1 are also numerical constants.

Proof. We have

$$\begin{aligned}
\sup_{\theta \in \mathcal{D}} \|K(\theta)\|_{\mathsf{op}} &= \sup_{\theta \in \mathcal{D}} \left\| \sum_{l=0}^{L-1} F_l(\theta)F_l^\top(\theta) \circ B_{l+1}(\theta)B_{l+1}^\top(\theta) \right\|_{\mathsf{op}} \\
&\leq \sum_{l=0}^{L-1} \sup_{\theta \in \mathcal{D}} \left\| F_l(\theta)F_l^\top(\theta) \circ B_{l+1}(\theta)B_{l+1}^\top(\theta) \right\|_{\mathsf{op}} \\
&\leq \sum_{l=0}^{L-1} \sup_{\theta \in \mathcal{D}} \left\| F_l(\theta)F_l^\top(\theta) \right\|_{\mathsf{op}} \sup_{\theta \in \mathcal{D}} \max_{i \in [n]} \|(B_{l+1}(\theta))_{i:}\|_2^2 \\
&\leq \sum_{l=0}^{L-1} \sup_{\theta \in \mathcal{D}} \|F_l(\theta)\|_F^2 \sup_{\theta \in \mathcal{D}} \max_{i \in [n]} \|b_{l+1}(\theta, x_i)\|_2^2 \\
&\leq \sum_{l=0}^{L-1} \left(\sum_{i=1}^n \sup_{\theta \in \mathcal{D}} \|f_l(\theta, x_i)\|_2^2 \right) \sup_{\theta \in \mathcal{D}} \max_i \|b_{l+1}(\theta, x_i)\|_2^2 .
\end{aligned} \tag{C.251}$$

By Lemma C.31, we have that $\sup_{\theta \in \mathcal{D}} \|f_l(\theta, x_i)\|_2^2 \leq C\,d$ with probability at least $1 - C\exp(-cN_{L-1})$ over $(W_k(\theta_0))_{k=1}^l$ and x_i, for any $0 \leq l \leq L-1$ and $i \in [n]$. By Lemma C.32, we have that $\sup_{\theta \in \mathcal{D}} \|b_l(\theta, x_i)\|_2^2 \leq C\gamma N_{L-1}$ with probability at least $1 - C\exp(-cN_{L-1})$ over $(W_k(\theta_0))_{k=l+1}^L$, for any $l \in [L]$ and $i \in [n]$. Therefore, we obtain

$$\sup_{\theta \in \mathcal{D}} \|K(\theta)\|_{\mathsf{op}} \leq C\,\gamma\,n\,d\,N_{L-1}, \tag{C.252}$$

with probability at least $1 - C\,n\exp(-cN_{L-1})$ over θ_0 and $(x_i)_{i=1}^n$, which gives the first statement of the lemma.

By using Weyl's inequality, we get

$$\begin{aligned}
\inf_{\theta \in \mathcal{D}} \sigma_{\mathsf{min}}(J(\theta)) &\geq \sigma_{\mathsf{min}}(J(\theta_0)) - \sup_{\theta \in \mathcal{D}} \|J(\theta_1) - J(\theta_0)\|_{\mathsf{op}} \\
&\geq c_1\sqrt{\gamma N_{L-2}N_{L-1}} - C_1\,R\,d\,\sqrt{\gamma\,N_{L-1}\,n},
\end{aligned} \tag{C.253}$$

where the last inequality follows from Lemma C.28 and Lemma C.34, and it holds with probability $1 - C\,ne^{-c\log^2 N_{L-1}} - Ce^{-c\log^2 n}$ over $(x_i)_{i=1}^n$ and θ_0. This gives the second statement of the lemma and concludes the proof. □

Armed with Proposition C.27 and the intermediate estimates of Lemmas C.28-C.35, we are finally ready to prove Theorem 2.11.

Proof of Theorem 2.11. We show that there exist two absolute constants $\tilde{c}$ and $\tilde{C}$ such that

$$\alpha = \tilde{c}\sqrt{\gamma N_{L-2} N_{L-1}} \tag{C.254}$$

and

$$\beta = \tilde{C}\sqrt{\gamma n d N_{L-1}} \tag{C.255}$$

satisfy the two assumptions in Proposition C.27 with initialization $\tilde{\theta}_0 := \theta_0$, where θ_0 is defined in (2.18). This holds with probability at least $1 - C\,ne^{-c\log^2 N_{L-1}} - Ce^{-c\log^2 n}$ over $(x_i)_{i=1}^n$ and θ_0.

Recall from Proposition C.27 that R is defined as $4\left\|F_L(\theta_0) - Y\right\|_2/\alpha$, since we have set $\tilde{\theta}_0 = \theta_0$. For the moment, we assume that

$$R = O\left(\sqrt{\frac{n}{\gamma N_{L-2} N_{L-1}}}\right), \tag{C.256}$$

and we will verify that this is the case later. Note that $\gamma = d^3 n^2 > 1$ and, hence, (C.256) and Assumption 2.5 imply that $R = o(1)$. Thus, we can apply Lemma C.35 and obtain

$$\inf_{\theta\in\mathcal{D}} \sigma_{\min}(J(\theta)) \geq c_1\sqrt{\gamma N_{L-2} N_{L-1}} - C_1\, R\, d\sqrt{\gamma N_{L-1} n} \geq \tilde{c}\sqrt{\gamma N_{L-2} N_{L-1}}, \tag{C.257}$$

with probability at least $1 - C\,ne^{-c\log^2 N_{L-1}} - Ce^{-c\log^2 n}$ over $(x_i)_{i=1}^n$ and θ_0, where the last inequality uses (C.256). This shows that the lower bound in (C.209) holds.

Now, by using (C.257), we verify that (C.256) holds. Recall that, by assumption of the theorem, $\|Y\|_2 = \Theta(\sqrt{n})$. Furthermore, by Lemma C.28, $F_L(\theta_0)$ is a vector of all zeros. Then,

$$R = \frac{4\left\|F_L(\theta_0) - Y\right\|_2}{\alpha} = \frac{4\|Y\|_2}{\alpha} = O\left(\sqrt{\frac{n}{\gamma N_{L-2} N_{L-1}}}\right). \tag{C.258}$$

By Lemma C.35, we have that

$$\sup_{\theta\in\mathcal{D}} \|J(\theta)\|_{\mathsf{op}} \leq C\sqrt{\gamma n d N_{L-1}}, \tag{C.259}$$

with probability at least $1 - C\,n\exp(-cN_{L-1})$ over θ_0. Thus, by our choice (C.255) of β, we obtain that the upper bound in (C.209) holds.

Next, we verify the second assumption of Proposition C.27. To do so, let us write

$$\frac{\alpha^2}{2\beta} = \frac{\tilde{c}^2 N_{L-2} N_{L-1}\gamma}{2\tilde{C}\sqrt{\gamma n d N_{L-1}}} = \Omega(\sqrt{N_{L-2}} n d), \tag{C.260}$$

where we have used Assumption 2.5. Thus,

$$\sup_{\theta_1,\theta_2\in\mathcal{D}} \|J(\theta_1) - J(\theta_2)\|_{\mathsf{op}} \leq C\, R\, d\sqrt{\gamma N_{L-1} n} = O\left(\frac{dn}{\sqrt{N_{L-2}}}\right) \leq \frac{\alpha^2}{2\beta}, \tag{C.261}$$

with probability at least $1 - C\,n\exp(-cN_{L-1})$ over $(x_i)_{i=1}^n$ and θ_0. Here, the first passage follows from Lemma C.34, in the second passage we use (C.256), and in the last one we use (C.260). This completes the proof of (C.210) and also of the theorem, since the desired claim follows from an application of Proposition C.27. □

APPENDIX D

Appendix for Chapter 3

D.1 Proof of Lemma 3.2

We start by refreshing some useful notions of linear algebra. Let $A \in \mathbb{R}^{n\times p}$ be a matrix, with $p \geq n$, and $A_{-1} \in \mathbb{R}^{(n-1)\times p}$ be obtained from A after removing the first row. We assume $AA^\top$ to be invertible, *i.e.,* the rows of A are linearly independent. Thus, also the rows of A_{-1} are linearly independent, implying that $A_{-1}A_{-1}^\top$ is invertible as well. We indicate with $P_A \in \mathbb{R}^{p\times p}$ the projector over $\mathrm{Span}\{\mathrm{rows}(A)\}$, and we correspondingly define $P_{A_{-1}} \in \mathbb{R}^{p\times p}$. As $AA^\top$ is invertible, we have that $\mathrm{rank}(A) = n$.

By singular value decomposition, we have $A = UDO^\top$, where $U \in \mathbb{R}^{n\times n}$ and $O \in \mathbb{R}^{p\times p}$ are orthogonal matrices, and $D \in \mathbb{R}^{n\times p}$ contains the (all strictly positive) singular values of A in its "left" diagonal, and is 0 in every other entry. Let us define $O_1 \in \mathbb{R}^{n\times p}$ as the matrix containing the first n rows of $O^\top$. This notation implies that if $O_1 u = 0$ for $u \in \mathbb{R}^p$, then $Au = 0$, *i.e.*, $u \in \mathrm{Span}\{\mathrm{rows}(A)\}^\perp$. The opposite implication is also true, which implies that $\mathrm{Span}\{\mathrm{rows}(A)\} = \mathrm{Span}\{\mathrm{rows}(O_1)\}$. As the rows of O_1 are orthogonal, we can then write

$$P_A = O_1^\top O_1. \tag{D.1}$$

We define $D_s \in \mathbb{R}^{n\times n}$, as the square, diagonal, and invertible matrix corresponding to the first n columns of D. Let's also define $I_n \in \mathbb{R}^{p\times p}$ as the matrix containing 1 in the first n entries of its diagonal, and 0 everywhere else. We have

$$\begin{aligned} P_A =& O_1^\top O_1 = O I_n O^\top \\ =& O D^\top D_s^{-2} D O^\top = O D^\top U^\top U D_s^{-2} U^\top U D O^\top \\ =& A^\top \left(U D_s^2 U^\top\right)^{-1} A = A^\top \left(U D O^\top O D^\top U^\top\right)^{-1} A \\ =& A^\top \left(A A^\top\right)^{-1} A \equiv A^+ A, \end{aligned} \tag{D.2}$$

where A^+ denotes the Moore-Penrose inverse.

Notice that this last form enables us to easily derive

$$P_{A_{-1}} A^+ v = A_{-1}^+ A_{-1} A^+ v = A_{-1}^+ I_{-1} A A^+ v = A_{-1}^+ I_{-1} v = A_{-1}^+ v_{-1}, \tag{D.3}$$

where $v \in \mathbb{R}^n$, $I_{-1} \in \mathbb{R}^{(n-1)\times n}$ is the $n \times n$ identity matrix without the first row, and $v_{-1} \in \mathbb{R}^{n-1}$ corresponds to v without its first entry.

Lemma D.1. *Let $\Phi \in \mathbb{R}^{n\times k}$ be a matrix whose first row is denoted as $\varphi(z_1)$. Let $\Phi_{-1} \in \mathbb{R}^{(n-1)\times k}$ be the original matrix without the first row, and let $P_{\Phi_{-1}}$ be the projector over the* span *of its rows. Then,*

$$\left\|P^{\perp}_{\Phi_{-1}}\varphi(z_1)\right\|_2^2 \geq \lambda_{\min}\left(\Phi\Phi^{\top}\right). \tag{D.4}$$

Proof. If $\lambda_{\min}\left(\Phi\Phi^{\top}\right) = 0$, the thesis becomes trivial. Otherwise, we have that $\Phi\Phi^{\top}$, and therefore $\Phi_{-1}\Phi_{-1}^{\top}$, are invertible.

Let $u \in \mathbb{R}^n$ be a vector, such that its first entry $u_1 = 1$. We denote with $u_{-1} \in \mathbb{R}^{n-1}$ the vector u without its first component, *i.e.* $u = [1, u_{-1}]$. We have

$$\left\|\Phi^{\top}u\right\|_2^2 \geq \lambda_{\min}\left(\Phi\Phi^{\top}\right)\|u\|_2^2 \geq \lambda_{\min}\left(\Phi\Phi^{\top}\right). \tag{D.5}$$

Setting $u_{-1} = -\left(\Phi_{-1}\Phi_{-1}^{\top}\right)^{-1}\Phi_{-1}\varphi(z_1)$, we get

$$\Phi^{\top}u = \varphi(z_1) + \Phi_{-1}^{\top}u_{-1} = \varphi(z_1) - P_{\Phi_{-1}}\varphi(z_1) = P^{\perp}_{\Phi_{-1}}\varphi(z_1). \tag{D.6}$$

Plugging this in (D.5), we get the thesis. □

At this point, we are ready to prove Lemma 3.2.

Proof of Lemma 3.2. We indicate with $\Phi_{-1} \in \mathbb{R}^{(n-1)\times p}$ the feature matrix of the training set $\Phi \in \mathbb{R}^{n\times p}$ *without* the first sample z_1. In other words, Φ_{-1} is equivalent to Φ, without the first row. Notice that since $K = \Phi\Phi^{\top}$ is invertible, also $K_{-1} := \Phi_{-1}\Phi_{-1}^{\top}$ is.

We can express the projector over the span of the rows of Φ in terms of the projector over the span of the rows of Φ_{-1} as follows

$$P_{\Phi} = P_{\Phi_{-1}} + \frac{P^{\perp}_{\Phi_{-1}}\varphi(z_1)\varphi(z_1)^{\top}P^{\perp}_{\Phi_{-1}}}{\left\|P^{\perp}_{\Phi_{-1}}\varphi(z_1)\right\|_2^2}. \tag{D.7}$$

The above expression is a consequence of the Gram-Schmidt formula, and the quantity at the denominator is different from zero because of Lemma D.1, as K is invertible.

We indicate with $\Phi^{+} = \Phi^{\top}K^{-1}$ the Moore-Penrose pseudo-inverse of Φ. Using (3.3), we can define $\theta^{*}_{-1} := \theta_0 + \Phi^{+}_{-1}\left(G_{-1} - f(Z_{-1}, \theta_0)\right)$, *i.e.*, the set of parameters the algorithm would have converged to if trained over (Z_{-1}, G_{-1}), the original data-set without the first pair sample-label (z_1, g_1).

Notice that $P_{\Phi}\Phi^{\top} = \Phi^{\top}$, as a consequence of (D.2). Thus, again using (3.3), for any z we can write

$$\begin{aligned} f(z,\theta^{*}) - \varphi(z)^{\top}\theta_0 &= \varphi(z)^{\top}\Phi^{+}\left(G - f(Z,\theta_0)\right) \\ &= \varphi(z)^{\top}P_{\Phi}\Phi^{+}\left(G - f(Z,\theta_0)\right) \\ &= \varphi(z)^{\top}\left(P_{\Phi_{-1}} + \frac{P^{\perp}_{\Phi_{-1}}\varphi(z_1)\varphi(z_1)^{\top}P^{\perp}_{\Phi_{-1}}}{\left\|P^{\perp}_{\Phi_{-1}}\varphi(z_1)\right\|_2^2}\right)\Phi^{+}\left(G - f(Z,\theta_0)\right). \end{aligned} \tag{D.8}$$

Notice that, thanks to (D.3), we can manipulate the first term in the bracket as follows

$$\begin{aligned} \varphi(z)^{\top}P_{\Phi_{-1}}\Phi^{+}\left(G - f(Z,\theta_0)\right) &= \varphi(z)^{\top}\Phi^{+}_{-1}\left(G_{-1} - f(Z_{-1},\theta_0)\right) \\ &= f(z,\theta^{*}_{-1}) - \varphi(z)^{\top}\theta_0. \end{aligned} \tag{D.9}$$

Thus, bringing the result of (D.9) on the LHS, (D.8) becomes

$$\begin{aligned} f(z,\theta^*) - f(z,\theta^*_{-1}) &= \frac{\varphi(z)^\top P^\perp_{\Phi_{-1}} \varphi(z_1)}{\left\| P^\perp_{\Phi_{-1}} \varphi(z_1) \right\|_2^2} \varphi(z_1)^\top P^\perp_{\Phi_{-1}} \Phi^+ \left(G - f(Z,\theta_0)\right) \\ &= \frac{\varphi(z)^\top P^\perp_{\Phi_{-1}} \varphi(z_1)}{\left\| P^\perp_{\Phi_{-1}} \varphi(z_1) \right\|_2^2} \varphi(z_1)^\top \left(I - P_{\Phi_{-1}}\right) \Phi^+ \left(G - f(Z,\theta_0)\right) \\ &= \frac{\varphi(z)^\top P^\perp_{\Phi_{-1}} \varphi(z_1)}{\left\| P^\perp_{\Phi_{-1}} \varphi(z_1) \right\|_2^2} \left(f(z_1,\theta^*) - f(z_1,\theta^*_{-1}) \right), \end{aligned} \tag{D.10}$$

where in the last step we again used (3.3) and (D.9). □

D.2 Useful lemmas

Lemma D.2. *Let x and y be two Lipschitz concentrated, independent random vectors. Let $\zeta(x,y)$ be a Lipschitz function in both arguments,* i.e., *for every δ,*

$$\begin{aligned} |\zeta(x+\delta,y) - \zeta(x,y)| &\le L \left\|\delta\right\|_2, \\ |\zeta(x,y+\delta) - \zeta(x,y)| &\le L \left\|\delta\right\|_2, \end{aligned} \tag{D.11}$$

for all x and y. Then, $\zeta(x,y)$ is a Lipschitz concentrated random variable, in the joint probability space of x and y.

Proof. To prove the thesis, we need to show that, for every 1-Lipschitz function τ, the following holds

$$\mathbb{P}_{xy}\left(\left|\tau\left(\zeta(x,y)\right) - \mathbb{E}_{xy}\left[\tau\left(\zeta(x,y)\right)\right]\right| > t\right) < e^{-ct^2}, \tag{D.12}$$

where c is a universal constant. An application of the triangle inequality gives

$$\begin{aligned} &\left|\tau\left(\zeta(x,y)\right) - \mathbb{E}_{xy}\left[\tau\left(\zeta(x,y)\right)\right]\right| \\ &\le \left|\tau\left(\zeta(x,y)\right) - \mathbb{E}_{x}\left[\tau\left(\zeta(x,y)\right)\right]\right| + \left|\mathbb{E}_{x}\left[\tau\left(\zeta(x,y)\right)\right] - \mathbb{E}_{y}\mathbb{E}_{x}\left[\tau\left(\zeta(x,y)\right)\right]\right| =: A + B. \end{aligned} \tag{D.13}$$

Thus, we can upper bound LHS of (D.12) as follows:

$$\mathbb{P}_{xy}\left(\left|\tau\left(\zeta(x,y)\right) - \mathbb{E}_{xy}\left[\tau\left(\zeta(x,y)\right)\right]\right| > t\right) \le \mathbb{P}_{xy}\left(A + B > t\right). \tag{D.14}$$

If A and B are positive random variables, it holds that $\mathbb{P}(A+B>t) \le \mathbb{P}(A>t/2) + \mathbb{P}(B>t/2)$. Then, the LHS of (D.12) is also upper bounded by

$$\mathbb{P}_{xy}\left(\left|\tau\left(\zeta(x,y)\right) - \mathbb{E}_{x}\left[\tau\left(\zeta(x,y)\right)\right]\right| > t/2\right) + \mathbb{P}_{xy}\left(\left|\mathbb{E}_{x}\left[\tau\left(\zeta(x,y)\right)\right] - \mathbb{E}_{y}\mathbb{E}_{x}\left[\tau\left(\zeta(x,y)\right)\right]\right| > t/2\right). \tag{D.15}$$

Since $\tau \circ \zeta$ is Lipschitz with respect to x for every y, we have

$$\mathbb{P}_{xy}\left(\left|\tau\left(\zeta(x,y)\right) - \mathbb{E}_{x}\left[\tau\left(\zeta(x,y)\right)\right]\right| > t/2\right) < e^{-c_1 t^2}, \tag{D.16}$$

for some absolute constant c_1. Furthermore, $\chi(y) := \mathbb{E}_x\left[\tau\left(\zeta(x,y)\right)\right]$ is also Lipschitz, as

$$\begin{aligned} \left|\chi(y+\delta) - \chi(y)\right| &= \left|\mathbb{E}_x\left[\tau\left(\zeta(x,y+\delta)\right) - \tau\left(\zeta(x,y)\right)\right]\right| \\ &\le \mathbb{E}_x\left[\left|\tau\left(\zeta(x,y+\delta)\right) - \tau\left(\zeta(x,y)\right)\right|\right] \\ &\le L\left\|\delta\right\|_2. \end{aligned} \tag{D.17}$$

Then, we can write

$$\mathbb{P}_{xy}\left(\left|\mathbb{E}_x\left[\tau\left(\zeta(x,y)\right)\right]-\mathbb{E}_y\mathbb{E}_x\left[\tau\left(\zeta(x,y)\right)\right]\right|>t/2\right)=\mathbb{P}_y\left(\left|\chi(y)-\mathbb{E}_y\left[\chi(y)\right]\right|>t/2\right)<e^{-c_2t^2}, \tag{D.18}$$

for some absolute constant c_2. Thus,

$$\mathbb{P}_{xy}\left(\left|\tau\left(\zeta(x,y)\right)-\mathbb{E}_{xy}\left[\tau\left(\zeta(x,y)\right)\right]\right|>t\right)<e^{-c_1t^2}+e^{-c_2t^2}\leq e^{-ct^2}, \tag{D.19}$$

for some absolute constant c, which concludes the proof. □

Lemma D.3. *Let $x\sim\mathcal{P}_X$, $y\sim\mathcal{P}_Y$ and $z=[x,y]\sim\mathcal{P}_Z$. Let Assumption 3.3 hold. Then, z is a Lipschitz concentrated random vector.*

Proof. We want to prove that, for every 1-Lipschitz function τ, the following holds

$$\mathbb{P}_z\left(\left|\tau\left(z\right)-\mathbb{E}_z\left[\tau\left(z\right)\right]\right|>t\right)<e^{-ct^2}, \tag{D.20}$$

for some universal constant c. As we can write $z=[x,y]$, defining $z'=[x',y]$, we have

$$\left|\tau\left(z\right)-\tau\left(z'\right)\right|\leq\left\|z-z'\right\|_2=\left\|x-x'\right\|_2, \tag{D.21}$$

i.e., for every y, τ is 1-Lipschitz with respect to x. The same can be shown for y, with an equivalent argument. Since x and y are independent random vectors, both Lipschitz concentrated, Lemma D.2 gives the thesis. □

Lemma D.4. *Let τ and ζ be two Lipschitz functions. Let $z,z'\in\mathbb{R}^d$ be two fixed vectors such that $\left\|z\right\|_2=\left\|z'\right\|_2=\sqrt{d}$. Let V be a $k\times d$ matrix such that $V_{i,j}\sim_{\text{i.i.d.}}\mathcal{N}(0,1/d)$. Then, for any $t>1$,*

$$\left|\tau(Vz)^\top\zeta(Vz')-\mathbb{E}_V\left[\tau(Vz)^\top\zeta(Vz')\right]\right|=O\left(\sqrt{k}\log t\right), \tag{D.22}$$

with probability at least $1-\exp(-c\log^2 t)$ over V. Here, τ and ζ act component-wise on their arguments. Furthermore, by taking $\tau=\zeta$ and $z=z'$, we have that

$$\mathbb{E}_V\left[\left\|\tau(Vz)\right\|_2^2\right]=k\mathbb{E}_\rho\left[\tau^2(\rho)\right], \tag{D.23}$$

where $\rho\sim\mathcal{N}(0,1)$. This implies that $\left\|\tau(Vz)\right\|_2^2=O\left(k\right)$ with probability at least $1-\exp(-ck)$ over V.

Proof. We have

$$\tau(Vz)^\top\zeta(Vz')=\sum_{j=1}^{k}\tau(v_j^\top z)\zeta(v_j^\top z'), \tag{D.24}$$

where we used the shorthand $v_j:=V_{j:}$. As τ and ζ are Lipschitz, $v_j\sim\mathcal{N}(0,I/d)$, and $\left\|z\right\|_2=\left\|z'\right\|_2=\sqrt{d}$, we have that $\tau(Vz)^\top\zeta(Vz')$ is the sum of k independent sub-exponential random variables, in the probability space of V. Thus, by Bernstein inequality (cf. Theorem 2.8.1 in Vershynin [2018]), we have

$$\left|\tau(Vz)^\top\zeta(Vz')-\mathbb{E}_V\left[\tau(Vz)^\top\zeta(Vz')\right]\right|=O\left(\sqrt{k}\log t\right). \tag{D.25}$$

with probability at least $1-\exp(-c\log^2 t)$, over the probability space of V, which gives the thesis. The second statement is again implied by the fact that $v_j\sim\mathcal{N}(0,I/d)$ and $\left\|z\right\|_2=\sqrt{d}$. □

Lemma D.5. *Let $x, x_1 \sim \mathcal{P}_X$ and $y_1 \sim \mathcal{P}_Y$ be independent random variables, with $x, x_1 \in \mathbb{R}^{d_x}$ and $y_1 \in \mathbb{R}^{d_y}$, and let Assumption 3.3 hold. Let $d = d_x + d_y$, V be a $k \times d$ matrix, such that $V_{i,j} \sim_{\text{i.i.d.}} \mathcal{N}(0, 1/d)$, and let τ be a Lipschitz function. Let $z_1 := [x_1, y_1]$ and $z_1^s := [x, y_1]$. Let $\alpha = d_y/d \in (0,1)$ and μ_l be the l-th Hermite coefficient of τ. Then, for any $t > 1$,*

$$\left| \tau(Vz_1^s)^\top \tau(Vz_1) - k \sum_{l=0}^{+\infty} \mu_l^2 \alpha^l \right| = O\left(\sqrt{k} \left(\sqrt{\frac{k}{d}} + 1 \right) \log t \right), \tag{D.26}$$

with probability at least $1 - \exp(-c \log^2 t) - \exp(-ck)$ over V and x, where c is a universal constant.

Proof. Define the vector x' as follows

$$x' = \frac{\sqrt{d_x} \left(I - \frac{x_1 x_1^\top}{d_x} \right) x}{\left\| \left(I - \frac{x_1 x_1^\top}{d_x} \right) x \right\|_2}. \tag{D.27}$$

Note that, by construction, $x_1^\top x' = 0$ and $\|x'\|_2 = \sqrt{d_x}$. Also, consider a vector y orthogonal to both x_1 and x. Then, a fast computation returns $y^\top x' = 0$. This means that x' is the vector on the $\sqrt{d_x}$-sphere, lying on the same plane of x_1 and x, orthogonal to x_1. Thus, we can easily compute

$$\frac{\left| x^\top x' \right|}{d_x} = \sqrt{1 - \left(\frac{x^\top x_1}{d_x} \right)^2} \geq 1 - \left(\frac{x^\top x_1}{d_x} \right)^2, \tag{D.28}$$

where the last inequality derives from $\sqrt{1-a} \geq 1 - a$ for $a \in [0,1]$. Then,

$$\|x - x'\|_2^2 = \|x\|_2^2 + \|x'\|_2^2 - 2x^\top x' \leq 2d_x \left(1 - \left(1 - \left(\frac{x^\top x_1}{d_x} \right)^2 \right) \right) = 2 \frac{\left(x^\top x_1 \right)^2}{d_x}. \tag{D.29}$$

As x and x_1 are both sub-Gaussian, mean-0 vectors, with ℓ_2 norm equal to $\sqrt{d_x}$, we have that

$$\mathbb{P}\left(\|x - x'\|_2 > t \right) \leq \mathbb{P}\left(|x^\top x_1| > \sqrt{d_x} t / \sqrt{2} \right) < \exp(-ct^2), \tag{D.30}$$

where c is an absolute constant. Here the probability is referred to the space of x, for a fixed x_1. Thus, $\|x - x'\|_2$ is sub-Gaussian.

We now define $z' := [x', y_1]$. Notice that $z_1^\top z' = \|y_1\|_2^2 = d_y$ and $\|z_1^s - z'\|_2 = \|x - x'\|_2$. We can write

$$\begin{aligned}
\left| \tau(Vz_1^s)^\top \tau(Vz_1) - \tau(Vz')^\top \tau(Vz_1) \right| &\leq \|\tau(Vz_1^s) - \tau(Vz')\|_2 \, \|\tau(Vz_1)\|_2 \\
&\leq C \|V\|_{\text{op}} \|z_1^s - z'\|_2 \, \|\tau(Vz_1)\|_2 \\
&\leq C_1 \left(\sqrt{\frac{k}{d}} + 1 \right) \|x - x'\|_2 \sqrt{k} \\
&= O\left(\sqrt{k} \left(\sqrt{\frac{k}{d}} + 1 \right) \log t \right).
\end{aligned} \tag{D.31}$$

Here the second step holds as τ is Lipschitz; the third step holds with probability at least $1-\exp(-c_1\log^2 t)-\exp(-c_2 k)$, and it uses rTheorem 4.4.5 of Vershynin [2018] and Lemma D.4; the fourth step holds with probability at least $1-\exp(-c\log^2 t)$, and it uses (D.30). This probability is intended over V and x. We further have

$$\left|\tau(Vz')^\top\tau(Vz_1)-\mathbb{E}_V\left[\tau(Vz')^\top\tau(Vz_1)\right]\right|=O\left(\sqrt{k}\log t\right), \tag{D.32}$$

with probability at least $1-\exp(-c_3\log^2 t)-\exp(-c_2 k)$ over V, because of Lemma D.4.

We have

$$\mathbb{E}_V\left[\tau(Vz')^\top\tau(Vz_1)\right]=k\mathbb{E}_{\rho_1\rho_2}\left[\tau(\rho_1)\tau(\rho_2)\right], \tag{D.33}$$

where we indicate with ρ_1 and ρ_2 two standard Gaussian random variables, with correlation

$$\mathsf{corr}(\rho_1,\rho_2)=\frac{z_1^\top z}{\|z_1\|_2\,\|z'\|_2}=\frac{d_y}{d}=\alpha. \tag{D.34}$$

Then, exploiting the Hermite expansion of τ, we have

$$\mathbb{E}_{\rho_1\rho_2}\left[\tau(\rho_1)\tau(\rho_2)\right]=\sum_{l=0}^{+\infty}\mu_l^2\alpha^l. \tag{D.35}$$

Putting together (D.31), (D.32), (D.33), and (D.35) gives the thesis. □

D.3 Proofs for random features

In this section, we indicate with $Z\in\mathbb{R}^{n\times d}$ the data matrix, such that its rows are sampled independently from $\mathcal{P}_Z$ (see Assumption 3.3). We denote by $V\in\mathbb{R}^{k\times d}$ the random features matrix, such that $V_{ij}\sim_{\text{i.i.d.}}\mathcal{N}(0,1/d)$. Thus, the feature map is given by (see (3.14))

$$\varphi(z):=\phi(Vz)\in\mathbb{R}^k, \tag{D.36}$$

where ϕ is the activation function, applied component-wise to the pre-activations Vz. We use the shorthands $\Phi:=\phi(ZV^\top)\in\mathbb{R}^{n\times k}$ and $K:=\Phi\Phi^\top\in\mathbb{R}^{n\times n}$, we indicate with $\Phi_{-1}\in\mathbb{R}^{(n-1)\times k}$ the matrix Φ without the first row, and we define $K_{-1}:=\Phi_{-1}\Phi_{-1}^\top$. We call P_Φ the projector over the span of the rows of Φ, and $P_{\Phi_{-1}}$ the projector over the span of the rows of Φ_{-1}. We use the notations $\tilde{\varphi}(z):=\varphi(z)-\mathbb{E}_V[\varphi(z)]$ and $\tilde{\Phi}_{-1}:=\Phi_{-1}-\mathbb{E}_V[\Phi_{-1}]$ to indicate the centered feature map and matrix respectively, where the centering is with respect to V. We indicate with μ_l the l-th Hermite coefficient of ϕ. We use the notation $z_1^s=[x,y_1]$, where $x\sim\mathcal{P}_X$ is sampled independently from V and Z. We denote by V_x (V_y) the first d_x (last d_y) columns of V, *i.e.*, $V=[V_x,V_y]$. We define $\alpha=d_y/d$. Throughout this section, for compactness, we drop the subscripts "RF" from these quantities, as we will only treat the proofs related to Section 3.5. Again for the sake of compactness, we will not re-introduce such quantities in the statements or the proofs of the following lemmas.

The content of this section can be summarized as follows:

- In Lemma D.7 we prove a lower bound on the smallest eigenvalue of K, adapting to our settings Lemma C.5 of Bombari et al. [2023]. As our assumptions are less restrictive than those in Bombari et al. [2023], we will crucially exploit Lemma D.6.

- In Lemma D.8, we treat separately a term that derives from $\mathbb{E}_V[\phi(Vz)] = \mu_0 \mathbf{1}_k$, showing that we can *center* the activation function, without changing our final statement in Theorem 3.6. This step is necessary only if $\mu_0 \neq 0$.

- In Lemma D.9, we show that the non-linear component of the features $\tilde{\varphi}(z_1) - \mu_1 V z_1$ and $\tilde{\varphi}(z_1^s) - \mu_1 V z_1^s$ have a negligible component in the space spanned by the rows of Φ_{-1}.

- In Lemma D.12, we provide concentration results for $\varphi(z_1^s)^\top P^\perp_{\Phi_{-1}} \varphi(z_1)$, and we lower bound this same term in Lemma D.11, exploiting also the intermediate result provided in Lemma D.10.

- Finally, we prove Theorem 3.6.

Lemma D.6. *Let $A := (Z^{*m}) \in \mathbb{R}^{n \times d^m}$, for some natural $m \geq 2$, where $*$ refers to the Khatri-Rao product, defined in Appendix B. We have*

$$\lambda_{\min}\left(AA^\top\right) = \Omega(d^m), \tag{D.37}$$

with probability at least $1 - \exp(-c \log^2 n)$ over Z, where c is an absolute constant.

Proof. As $m \geq 2$, we can write $A = (Z^{*2}) * \left(Z^{*(m-2)}\right) =: A_2 * A_m$ (where (Z^{*0}) is defined to be the vector full of ones $\mathbf{1}_n \in \mathbb{R}^n$). We can provide a lower bound on the smallest eigenvalue of such product through the following inequality Schur [1911]:

$$\lambda_{\min}\left(AA^\top\right) = \lambda_{\min}\left(A_2 A_2^\top \circ A_m A_m^\top\right) \geq \lambda_{\min}\left(A_2 A_2^\top\right) \min_i \|(A_m)_{i:}\|_2^2 . \tag{D.38}$$

Note that the rows of Z are mean-0 and Lipschitz concentrated by Lemma D.3. Then, by following the argument of Lemma C.3 in Bombari et al. [2023], we have

$$\lambda_{\min}\left(A_2 A_2^\top\right) = \Omega(d^2), \tag{D.39}$$

with probability at least $1 - \exp(-c \log^2 n)$ over Z. We remark that, for the argument of Lemma C.3 in Bombari et al. [2023] to go through, it suffices that $n = o(d^2/\log^4 d)$ and $n \log^4 n = o(d^2)$ (see Equations (C.23) and (C.26) in Bombari et al. [2023]), which is implied by Assumption 3.4, despite it being milder than Assumption 4 in Bombari et al. [2023].

For the second term of (D.38), we have

$$\|(A_m)_{i:}\|_2^2 = \|z_i\|_2^{2(m-2)} = d^{m-2}, \tag{D.40}$$

due to Assumption 3.3. Thus, the thesis readily follows. □

Lemma D.7. *We have that*

$$\lambda_{\min}(K) = \Omega(k), \tag{D.41}$$

with probability at least $1 - \exp\left(-c \log^2 n\right)$ over V and Z, where c is an absolute constant. This implies that $\lambda_{\min}(K_{-1}) = \Omega(k)$.

Proof. The proof follows the same path as Lemma C.5 of Bombari et al. [2023]. In particular, we define a truncated version of Φ as follows

$$\bar{\Phi}_{:j} = \phi(Zv_j)\chi\left(\|\phi(Zv_j)\|_2^2 \leq R\right), \tag{D.42}$$

where χ is the indicator function and we introduce the shorthand $v_i := V_{i:}$. In this case, $\chi = 1$ if $\|\phi(Zv_j)\|_2^2 \leq R$, and $\chi = 0$ otherwise. As this is a column-wise truncation, it's easy to verify that $\Phi\Phi^\top \succeq \bar{\Phi}\bar{\Phi}^\top$. Over such truncated matrix, we can use Matrix Chernoff inequality (see Theorem 1.1 of Tropp [2012]), which gives that $\lambda_{\min}\left(\bar{\Phi}\bar{\Phi}^\top\right) = \Omega(\lambda_{\min}\left(\bar{G}\right))$, where $\bar{G} := \mathbb{E}_V\left[\bar{\Phi}\bar{\Phi}^\top\right]$. Finally, we prove closeness between $\bar{G}$ and G, which is analogously defined as $G := \mathbb{E}_V\left[\Phi\Phi^\top\right]$.

To be more specific, setting $R = k/\log^2 n$, we have

$$\lambda_{\min}(K) \geq \lambda_{\min}\left(\bar{\Phi}\bar{\Phi}^\top\right) \geq \lambda_{\min}\left(\bar{G}\right)/2 \geq \lambda_{\min}(G)/2 - o(k), \tag{D.43}$$

where the second inequality holds with probability at least $1 - \exp(c_1 \log^2 n)$ over V, if $\lambda_{\min}(G) = \Omega(k)$ (see Equation (C.47) of Bombari et al. [2023]), and the third comes from Equation (C.45) in Bombari et al. [2023]. To perform these steps, our Assumptions 3.4 and 3.5 are enough, despite the second one being milder than Assumption 2 in Bombari et al. [2023].

To conclude the proof, we are left to prove that $\lambda_{\min}(G) = \Omega(k)$ with probability at least $1 - \exp(-c_2 \log^2 n)$ over V and Z.

We have that

$$G = \mathbb{E}_V[K] = \mathbb{E}_V\left[\sum_{i=1}^{k} \phi(ZV_{i:}^\top)\phi(ZV_{i:}^\top)^\top\right] = k\mathbb{E}_v\left[\phi(Zv)\phi(Zv)^\top\right] := kM, \tag{D.44}$$

where we use the shorthand v to indicate a random variable distributed as $V_{1:}$. We also indicate with z_i the i-th row of Z. Exploiting the Hermite expansion of ϕ, we can write

$$M_{ij} = \mathbb{E}_v\left[\phi(z_i^\top v)\phi(z_j^\top v)\right] = \sum_{l=0}^{+\infty} \mu_l^2 \frac{\left(z_i^\top z_j\right)^l}{d^l} = \sum_{l=0}^{+\infty} \mu_l^2 \frac{\left[\left(Z^{*l}\right)\left(Z^{*l}\right)^\top\right]_{ij}}{d^l}, \tag{D.45}$$

where μ_l is the l-th Hermite coefficient of ϕ. Note that the previous expansion was possible since $\|z_i\| = \sqrt{d}$ for all $i \in [n]$. As ϕ is non-linear, there exists $m \geq 2$ such that $\mu_m^2 > 0$. In particular, we have $M \succeq \frac{\mu_m^2}{d^m} AA^\top$ in a PSD sense, where we define

$$A := \left(Z^{*m}\right). \tag{D.46}$$

By Lemma D.6, the desired result readily follows. □

Lemma D.8. *Let $\mu_0 \neq 0$. Then,*

$$\left\|P_{\Phi_{-1}}^{\perp}\mathbf{1}_k\right\|_2 = o(\sqrt{k}), \tag{D.47}$$

with probability at least $1 - e^{-cd} - e^{-cn}$ over V and Z, where c is an absolute constant.

Proof. Note that $\Phi_{-1}^\top = \mu_0 \mathbf{1}_k \mathbf{1}_{n-1}^\top + \tilde{\Phi}_{-1}^\top$. Here, $\tilde{\Phi}_{-1}^\top$ is a $k \times (n-1)$ matrix with i.i.d. and mean-0 rows, whose sub-Gaussian norm (in the probability space of V) can be bounded as

$$\left\|\tilde{\Phi}_{:i}\right\|_{\psi_2} = \|\phi(ZV_{i:}) - \mathbb{E}_V[\phi(ZV_{i:})]\|_{\psi_2} \leq L\frac{\|Z\|_{\mathsf{op}}}{\sqrt{d}} = O\left(\sqrt{n/d}+1\right), \tag{D.48}$$

where first inequality holds since ϕ is L-Lipschitz and $V_{i:}$ is a Gaussian (and hence, Lipschitz concentrated) vector with covariance I/d. The last step holds with probability at least $1-e^{-cd}$ over Z, because of Lemma B.7 in Bombari et al. [2022b].

Thus, another application of Lemma B.7 in Bombari et al. [2022b] gives

$$\left\|\tilde{\Phi}_{-1}^\top\right\|_{\mathsf{op}} = O\left(\left(\sqrt{k}+\sqrt{n}\right)\left(\sqrt{n/d}+1\right)\right) = O\left(\sqrt{k}\left(\sqrt{n/d}+1\right)\right), \tag{D.49}$$

where the first equality holds with probability at least $1-e^{-cn}$ over V, and the second is a direct consequence of Assumption 3.4.

We can write

$$\Phi_{-1}^\top \frac{\mathbf{1}_{n-1}}{\mu_0(n-1)} = \left(\mu_0 \mathbf{1}_k \mathbf{1}_{n-1}^\top + \tilde{\Phi}_{-1}^\top\right)\frac{\mathbf{1}_{n-1}}{\mu_0(n-1)} = \mathbf{1}_k + \tilde{\Phi}_{-1}^\top \frac{\mathbf{1}_{n-1}}{\mu_0(n-1)} =: \mathbf{1}_k + v, \tag{D.50}$$

where

$$\|v\|_2 \leq \frac{1}{\mu_0(n-1)}\left\|\tilde{\Phi}_{-1}^\top\right\|_{\mathsf{op}}\|\mathbf{1}_{n-1}\|_2 = O\left(\sqrt{\frac{k}{n}}\left(\sqrt{n/d}+1\right)\right) = o(\sqrt{k}). \tag{D.51}$$

Thus, we can conclude

$$\begin{aligned}\left\|P_{\Phi_{-1}}^\perp \mathbf{1}_k\right\|_2 &= \left\|P_{\Phi_{-1}}^\perp\left(\Phi_{-1}^\top \frac{\mathbf{1}_{n-1}}{\mu_0(n-1)} - v\right)\right\|_2 \\ &\leq \left\|P_{\Phi_{-1}}^\perp P_{\Phi_{-1}} \Phi_{-1}^\top \frac{\mathbf{1}_{n-1}}{\mu_0(n-1)}\right\|_2 + \|v\|_2 = o(\sqrt{k}),\end{aligned} \tag{D.52}$$

where in the second step we use the triangle inequality, $\Phi_{-1}^\top = P_{\Phi_{-1}}\Phi_{-1}^\top$, and $\left\|P_{\Phi_{-1}}^\perp v\right\|_2 \leq \|v\|_2$. □

Lemma D.9. *Let $z \sim \mathcal{P}_Z$, sampled independently from Z_{-1}, and denote $\tilde{\phi}(x) := \phi(x) - \mu_0$. Then,*

$$\left\|P_{\Phi_{-1}}\left(\tilde{\phi}(Vz) - \mu_1 Vz\right)\right\|_2 = o(\sqrt{k}), \tag{D.53}$$

with probability at least $1 - \exp\left(-c\log^2 n\right)$ over V, Z_{-1} and z, where c is an absolute constant.

Proof. As $P_{\Phi_{-1}} = \Phi_{-1}^+ \Phi_{-1}$, we have

$$\begin{aligned}\left\|P_{\Phi_{-1}}\left(\tilde{\phi}(Vz) - \mu_1 Vz\right)\right\|_2 &\leq \left\|\Phi_{-1}^+\right\|_{\mathsf{op}}\left\|\Phi_{-1}\left(\tilde{\phi}(Vz) - \mu_1 Vz\right)\right\|_2 \\ &= O\left(\frac{\left\|\Phi_{-1}\left(\tilde{\phi}(Vz) - \mu_1 Vz\right)\right\|_2}{\sqrt{k}}\right),\end{aligned} \tag{D.54}$$

where the last equality holds with probability at least $1 - \exp\left(-c\log^2 n\right)$ over V and Z_{-1}, because of Lemma D.7.

An application of Lemma D.4 with $t = n$ gives

$$|u_i - \mathbb{E}_V[u_i]| = O\left(\sqrt{k}\log n\right), \tag{D.55}$$

where u_i is the i-th entry of the vector $u := \Phi_{-1}\left(\tilde{\phi}(Vz) - \mu_1 Vz\right)$. This can be done since both ϕ and $\tilde{\phi} \equiv \phi - \mu_0$ are Lipschitz, $v_j \sim \mathcal{N}(0, I/d)$, and $\|z\|_2 = \|z_{i+1}\|_2 = \sqrt{d}$. Performing a union bound over all entries of u, we can guarantee that the previous equation holds for every $1 \le i \le n-1$, with probability at least $1 - (n-1)\exp(-c\log^2 n) \ge 1 - \exp(-c_1\log^2 n)$. Thus, we have

$$\|u - \mathbb{E}_V[u]\|_2 = O\left(\sqrt{k}\sqrt{n}\log n\right) = o(k), \tag{D.56}$$

where the last equality holds because of Assumption 3.4.

Note that the function $f(x) := \tilde{\phi}(x) - \mu_1 x$ has the first 2 Hermite coefficients equal to 0. Hence, as $v_i^\top z$ and $v_i^\top z_i$ are standard Gaussian random variables with correlation $\frac{z^\top z_i}{\|z\|_2\|z_i\|_2}$, we have

$$\begin{aligned}
|\mathbb{E}_V[u_i]| &\le k\sum_{l=2}^{+\infty}\mu_l^2\left(\frac{\left|z^\top z_i\right|}{\|z\|_2\|z_i\|_2}\right)^l \\
&\le k\max_l\mu_l^2\sum_{l=2}^{+\infty}\left(\frac{\left|z^\top z_i\right|}{\|z\|_2\|z_i\|_2}\right)^l \\
&= k\max_l\mu_l^2\left(\frac{z^\top z_i}{\|z\|_2\|z_i\|_2}\right)^2\frac{1}{1-\frac{\left|z^\top z_i\right|}{\|z\|_2\|z_i\|_2}} \\
&\le 2k\max_l\mu_l^2\left(\frac{z^\top z_i}{\|z\|_2\|z_i\|_2}\right)^2 = O\left(\frac{k\log^2 n}{d}\right),
\end{aligned} \tag{D.57}$$

where the last inequality holds with probability at least $1 - \exp\left(-c\log^2 n\right)$ over z and z_i, as they are two independent, mean-0, sub-Gaussian random vectors. Again, performing a union bound over all entries of $\mathbb{E}_V[u]$, we can guarantee that the previous equation holds for every $1 \le i \le n-1$, with probability at least $1 - (n-1)\exp(-c\log^2 n) \ge 1 - \exp(-c_1\log^2 n)$. Then, we have

$$\|\mathbb{E}_V[u]\|_2 = O\left(\sqrt{n}\,\frac{k\log^2 n}{d}\right) = o(k), \tag{D.58}$$

where the last equality is a consequence of Assumption 3.4.

Finally, (D.56) and (D.58) give

$$\left\|\Phi_{-1}\left(\tilde{\phi}(Vz) - \mu_1 Vz\right)\right\|_2 \le \|\mathbb{E}_V[u]\|_2 + \|u - \mathbb{E}_V[u]\|_2 = o(k), \tag{D.59}$$

which plugged in (D.54) readily provides the thesis. □

Lemma D.10. *We have*

$$\left|\left(Vz_1^s\right)^\top P^\perp_{\Phi_{-1}} Vz_1 - \left\|P^\perp_{\Phi_{-1}} V_y y_1\right\|_2^2\right| = o(k), \tag{D.60}$$

with probability at least $1 - \exp(-c\log^2 n)$ *over* x, z_1 *and* V, *where* c *is an absolute constant.*

Proof. We have

$$Vz_1^s = V_x x + V_y y_1, \qquad Vz_1 = V_x x_1 + V_y y_1. \tag{D.61}$$

Thus, we can write

$$\begin{aligned}\left|(Vz_1^s)^\top P_{\Phi_{-1}}^\perp Vz_1 - \left\|P_{\Phi_{-1}}^\perp V_y y_1\right\|_2^2\right| &= \left|(V_x x)^\top P_{\Phi_{-1}}^\perp Vz_1 + (V_y y_1)^\top P_{\Phi_{-1}}^\perp V_x x_1\right| \\ &\le \left|x^\top V_x^\top P_{\Phi_{-1}}^\perp Vz_1\right| + \left|y_1^\top V_y^\top P_{\Phi_{-1}}^\perp V_x x_1\right|.\end{aligned} \tag{D.62}$$

Let's look at the first term of the RHS of the previous equation. Notice that $\|V\|_{\mathsf{op}} = O\left(\sqrt{k/d}+1\right)$ with probability at least $1-2e^{-cd}$, because of Theorem 4.4.5 of Vershynin [2018]. We condition on such event until the end of the proof, which also implies having the same bound on $\|V_x\|_{\mathsf{op}}$ and $\|V_y\|_{\mathsf{op}}$. Since x is a mean-0 sub-Gaussian vector, independent from $V_x^\top P_{\Phi_{-1}}^\perp Vz_1$, we have

$$\begin{aligned}\left|x^\top V_x^\top P_{\Phi_{-1}}^\perp Vz_1\right| &\le \log n \left\|V_x^\top P_{\Phi_{-1}}^\perp Vz_1\right\|_2 \\ &\le \log n \|V_x\|_{\mathsf{op}} \left\|P_{\Phi_{-1}}^\perp\right\|_{\mathsf{op}} \|V\|_{\mathsf{op}} \|z_1\| \\ &= O\left(\log n \left(\frac{k}{d}+1\right)\sqrt{d}\right) = o(k),\end{aligned} \tag{D.63}$$

where the first inequality holds with probability at least $1-\exp(-c\log^2 n)$ over x, and the last line holds because $\left\|P_{\Phi_{-1}}^\perp\right\|_{\mathsf{op}} \le 1$, $\|z_1\| = \sqrt{d}$, and because of Assumption 3.4.

Similarly, exploiting the independence between x_1 and y_1, we can prove that $\left|y_1^\top V_y^\top P_{\Phi_{-1}}^\perp V_x x_1\right| = o(k)$, with probability at least $1-\exp(-c\log^2 n)$ over y_1. Plugging this and (D.63) in (D.62) readily gives the thesis. □

Lemma D.11. *We have*

$$\left|\varphi(z_1^s)^\top P_{\Phi_{-1}}^\perp \varphi(z_1) - \left(k\left(\sum_{l=2}^{+\infty}\mu_l^2\alpha^l\right) + \mu_1^2\left\|P_{\Phi_{-1}}^\perp V_y y_1\right\|_2^2\right)\right| = o(k), \tag{D.64}$$

with probability at least $1-\exp(-c\log^2 n)$ *over* V *and* Z*, where* c *is an absolute constant.*

Proof. An application of Lemma D.4 and Assumption 3.4 gives

$$\begin{aligned}\|\varphi(z_1)\|_2 &= O\left(\sqrt{k}\right), \qquad \|\varphi(z_1^s)\|_2 = O\left(\sqrt{k}\right), \\ \|Vz_1\|_2 &= O\left(\sqrt{k}\right), \qquad \|Vz_1^s\|_2 = O\left(\sqrt{k}\right),\end{aligned} \tag{D.65}$$

with probability at least $1-\exp(-c_1\log^2 n)$ over V, where c_1 is an absolute constant. We condition on such high probability event until the end of the proof.

Let's suppose $\mu_0 \neq 0$. Then, we have

$$\left|\varphi(z_1^s)^\top P_{\Phi_{-1}}^\perp \varphi(z_1) - \tilde{\phi}(Vz_1^s)^\top P_{\Phi_{-1}}^\perp \tilde{\phi}(Vz_1)\right| = o(k), \tag{D.66}$$

with probability at least $1-\exp(c_2\log^2 n)$ over V and Z, because of (D.65) and Lemma D.8. Note that (D.66) trivially holds even when $\mu_0 = 0$, as $\phi \equiv \tilde{\phi}$. Thus, (D.66) is true in any case with probability at least $1-\exp(c_2\log^2 n)$ over V and Z.

Furthermore, because of (D.65) and Lemma D.9, we have

$$\left|\tilde{\phi}(Vz_1^s)^\top P_{\Phi_{-1}}\tilde{\phi}(Vz_1) - \mu_1^2(Vz_1^s)^\top P_{\Phi_{-1}}(Vz_1)\right| = o(k), \tag{D.67}$$

with probability at least $1-\exp(-c_3\log^2 n)$ over V and Z.

Thus, putting (D.66) and (D.67) together, and using Lemma D.10, we get

$$\begin{aligned}
&\left|\varphi(z_1^s)^\top P^\perp_{\Phi_{-1}}\varphi(z_1) - \left(\tilde{\phi}(Vz_1^s)^\top\tilde{\phi}(Vz_1) - \mu_1^2(Vz_1^s)^\top(Vz_1) + \mu_1^2\left\|P^\perp_{\Phi_{-1}}V_y y_1\right\|_2^2\right)\right| \\
&\quad\le \left|\varphi(z_1^s)^\top P^\perp_{\Phi_{-1}}\varphi(z_1) - \tilde{\phi}(Vz_1^s)^\top P^\perp_{\Phi_{-1}}\tilde{\phi}(Vz_1)\right| \\
&\qquad + \left|-\tilde{\phi}(Vz_1^s)^\top P_{\Phi_{-1}}\tilde{\phi}(Vz_1) + \mu_1^2(Vz_1^s)^\top P_{\Phi_{-1}}(Vz_1)\right| \\
&\qquad + \left|\mu_1^2(Vz_1^s)^\top P^\perp_{\Phi_{-1}}(Vz_1) - \mu_1^2\left\|P^\perp_{\Phi_{-1}}V_y y_1\right\|_2^2\right| = o(k),
\end{aligned} \tag{D.68}$$

with probability at least $1-\exp(-c_4\log^2 n)$ over V and X and x. To conclude we apply Lemma D.5 setting $t=n$, together with Assumption 3.4, to get

$$\left|\tilde{\phi}(Vz_1^s)^\top\tilde{\phi}(Vz_1) - k\left(\sum_{l=1}^{+\infty}\mu_l^2\alpha^l\right)\right| = O\left(\sqrt{k}\left(\sqrt{\frac{k}{d}}+1\right)\log n\right) = o(k), \tag{D.69}$$

and

$$\left|\mu_1^2(Vz_1^s)^\top(Vz_1) - k\mu_1^2\alpha\right| = O\left(\sqrt{k}\left(\sqrt{\frac{k}{d}}+1\right)\log n\right) = o(k), \tag{D.70}$$

which jointly hold with probability at least $1-\exp(-c_5\log^2 n)$ over V and x.

Applying the triangle inequality to (D.68), (D.69), and (D.70), we get the thesis. □

Lemma D.12. *We have that*

$$\left|\left\|P^\perp_{\Phi_{-1}}\varphi(z_1)\right\|_2^2 - \mathbb{E}_{z_1}\left[\left\|P^\perp_{\Phi_{-1}}\varphi(z_1)\right\|_2^2\right]\right| = o(k), \tag{D.71}$$

$$\left|\varphi(z_1^s)^\top P^\perp_{\Phi_{-1}}\varphi(z_1) - \mathbb{E}_{z_1,z_1^s}\left[\varphi(z_1^s)^\top P^\perp_{\Phi_{-1}}\varphi(z_1)\right]\right| = o(k), \tag{D.72}$$

jointly hold with probability at least $1-\exp(-c\log^2 n)$ over z_1, V and x, where c is an absolute constant.

Proof. Let's condition until the end of the proof on both $\|V_x\|_{\mathsf{op}}$ and $\|V_y\|_{\mathsf{op}}$ to be $O\left(\sqrt{k/d}+1\right)$, which happens with probability at least $1-e^{-c_1 d}$ by Theorem 4.4.5 of Vershynin [2018]. This also implies that $\|V\|_{\mathsf{op}} = O\left(\sqrt{k/d}+1\right)$.

We indicate with $\nu := \mathbb{E}_{z_1}[\varphi(z_1)] = \mathbb{E}_{z_1^s}[\varphi(z_1^s)] \in \mathbb{R}^k$, and with $\hat{\varphi}(z) := \varphi(z) - \nu$. Note that, as φ is a $C\left(\sqrt{k/d}+1\right)$-Lipschitz function, for some constant C, and as z_1 is Lipschitz concentrated, by Assumption3.4, we have

$$\left|\|\varphi(z_1)\|_2 - \mathbb{E}_{z_1}[\|\varphi(z_1)\|_2]\right| = o\left(\sqrt{k}\right), \tag{D.73}$$

with probability at least $1-\exp(-c_2\log^2 n)$ over z_1 and V. In addition, by the last statement of Lemma D.4 and Assumption 3.4, we have that $\|\varphi(z_1)\|_2 = O\left(\sqrt{k}\right)$ with probability $1-\exp(-c_3\log^2 n)$ over V. Thus, taking the intersection between these two events, we have

$$\mathbb{E}_{z_1}[\|\varphi(z_1)\|_2] = O\left(\sqrt{k}\right), \tag{D.74}$$

with probability at least $1 - \exp(-c_4 \log^2 n)$ over z_1 and V. As this statement is independent of z_1, it holds with the same probability just over the probability space of V. Then, by Jensen inequality, we have

$$\|\nu\|_2 = \|\mathbb{E}_{z_1}[\varphi(z_1)]\|_2 \leq \mathbb{E}_{z_1}[\|\varphi(z_1)\|_2] = O\left(\sqrt{k}\right). \tag{D.75}$$

We can now rewrite the LHS of the first statement as

$$\begin{aligned}
&\left| \left\| P^{\perp}_{\Phi_{-1}} \varphi(z_1) \right\|_2^2 - \mathbb{E}_{z_1}\left[\left\| P^{\perp}_{\Phi_{-1}} \varphi(z_1) \right\|_2^2 \right] \right| \\
&= \left| \left\| P^{\perp}_{\Phi_{-1}} \left(\hat{\varphi}(z_1) + \nu \right) \right\|_2^2 - \mathbb{E}_{z_1}\left[\left\| P^{\perp}_{\Phi_{-1}} \left(\hat{\varphi}(z_1) + \nu \right) \right\|_2^2 \right] \right| \\
&= \left| \hat{\varphi}(z_1)^\top P^{\perp}_{\Phi_{-1}} \hat{\varphi}(z_1) + 2\nu^\top P^{\perp}_{\Phi_{-1}} \hat{\varphi}(z_1) - \mathbb{E}_{z_1}\left[\hat{\varphi}(z_1)^\top P^{\perp}_{\Phi_{-1}} \hat{\varphi}(z_1) \right] \right| \\
&\leq \left| \hat{\varphi}(z_1)^\top P^{\perp}_{\Phi_{-1}} \hat{\varphi}(z_1) - \mathbb{E}_{z_1}\left[\hat{\varphi}(z_1)^\top P^{\perp}_{\Phi_{-1}} \hat{\varphi}(z_1) \right] \right| + 2 \left| \nu^\top P^{\perp}_{\Phi_{-1}} \hat{\varphi}(z_1) \right|.
\end{aligned} \tag{D.76}$$

The second term is the inner product between $\hat{\varphi}(z_1)$, a mean-0 sub-Gaussian vector (in the probability space of z_1) such that $\|\hat{\varphi}(z_1)\|_{\psi_2} = O\left(\sqrt{k/d} + 1\right)$, and the independent vector $P^{\perp}_{\Phi_{-1}} \nu$, such that $\left\| P^{\perp}_{\Phi_{-1}} \nu \right\|_2 \leq \|\nu\|_2 = O\left(\sqrt{k}\right)$, because of (D.75). Thus, by Assumption3.4, we have that

$$\left| \nu^\top P^{\perp}_{\Phi_{-1}} \hat{\varphi}(z_1) \right| = o(k), \tag{D.77}$$

with probability at least $1 - \exp(-c_5 \log^2 n)$ over z_1 and V. Then, as $\left(\sqrt{k/d} + 1\right)^{-1} \hat{\varphi}(z_1)$ is a mean-0, Lipschitz concentrated random vector (in the probability space of z_1), by the general version of the Hanson-Wright inequality given by Theorem 2.3 in Adamczak [2015], we can write

$$\begin{aligned}
&\mathbb{P}\left(\left| \left\| P^{\perp}_{\Phi_{-1}} \hat{\varphi}(z_1) \right\|_2^2 - \mathbb{E}_{z_1}\left[\left\| P^{\perp}_{\Phi_{-1}} \hat{\varphi}(z_1) \right\|_2^2 \right] \right| \geq k / \log n \right) \\
&\leq 2 \exp\left(-c_6 \min\left(\frac{k^2}{\log^2 n \left((k/d)^2 + 1\right) \left\| P^{\perp}_{\Phi_{-1}} \right\|_F^2}, \frac{k}{\log n \left(k/d + 1\right) \left\| P^{\perp}_{\Phi_{-1}} \right\|_{\mathsf{op}}} \right) \right) \\
&\leq 2 \exp\left(-c_6 \min\left(\frac{k}{\log^2 n \left((k/d)^2 + 1\right)}, \frac{k}{\log n \left(k/d + 1\right)} \right) \right) \\
&\leq \exp\left(-c_7 \log^2 n \right),
\end{aligned} \tag{D.78}$$

where the last inequality comes from Assumption 3.4.This, together with (D.76) and (D.77), proves the first part of the statement.

For the second part of the statement, we have

$$\begin{aligned}
&\left| \varphi(z_1^s)^\top P^{\perp}_{\Phi_{-1}} \varphi(z_1) - \mathbb{E}_{z_1, z_1^s}\left[\varphi(z_1^s)^\top P^{\perp}_{\Phi_{-1}} \varphi(z_1) \right] \right| \\
&\leq \left| \hat{\varphi}(z_1^s)^\top P^{\perp}_{\Phi_{-1}} \hat{\varphi}(z_1) - \mathbb{E}_{z_1, z_1^s}\left[\hat{\varphi}(z_1^s)^\top P^{\perp}_{\Phi_{-1}} \hat{\varphi}(z_1) \right] \right| + \left| \nu^\top P^{\perp}_{\Phi_{-1}} \hat{\varphi}(z_1) \right| + \left| \nu^\top P^{\perp}_{\Phi_{-1}} \hat{\varphi}(z_1^s) \right|.
\end{aligned} \tag{D.79}$$

Following the same argument that led to (D.77), we obtain

$$\left| \nu^\top P^{\perp}_{\Phi_{-1}} \hat{\varphi}(z_1^s) \right| = o(k), \tag{D.80}$$

with probability at least $1 - \exp(-c_8 \log^2 n)$ over z_1^s and V. Let us set

$$P_2 := \frac{1}{2} \left(\begin{array}{c|c} 0 & P^{\perp}_{\Phi_{-1}} \\ \hline P^{\perp}_{\Phi_{-1}} & 0 \end{array} \right), \qquad V_2 := \left(\begin{array}{c|c|c} V_x & V_y & 0 \\ \hline 0 & V_y & V_x \end{array} \right), \tag{D.81}$$

and

$$\hat{\varphi}_2 := \phi\left(V_2[x_1, y_1, x]^\top\right) - \mathbb{E}_{x_1,y_1,x}\left[\phi\left(V_2[x_1, y_1, x]^\top\right)\right] \equiv [\hat{\varphi}(z_1), \hat{\varphi}(z_1^s)]^\top. \tag{D.82}$$

We have that $\|P_2\|_{\mathsf{op}} \leq 1$, $\|P_2\|_F^2 \leq k$, $\|V_2\|_{\mathsf{op}} \leq 2\|V_x\|_{\mathsf{op}} + 2\|V_y\|_{\mathsf{op}} = O\left(\sqrt{k/d}+1\right)$, and that $[x_1, y_1, x]^\top$ is a Lipschitz concentrated random vector in the joint probability space of z_1 and z_1^s, which follows from applying Lemma D.3 twice. Also, we have

$$\hat{\varphi}(z_1^s)^\top P_{\Phi_{-1}}^\perp \hat{\varphi}(z_1) = \hat{\varphi}_2^\top P_2 \hat{\varphi}_2. \tag{D.83}$$

Thus, as $\left(\sqrt{k/d}+1\right)^{-1}\hat{\varphi}_2$ is a mean-0, Lipschitz concentrated random vector (in the probability space of z_1 and z_1^s), again by the general version of the Hanson-Wright inequality given by Theorem 2.3 in Adamczak [2015], we can write

$$\begin{aligned}
&\mathbb{P}\left(\left|\hat{\varphi}_2^\top P_2 \hat{\varphi}_2 - \mathbb{E}_{z_1,z_1^s}\left[\hat{\varphi}_2^\top P_2 \hat{\varphi}_2\right]\right| \geq k/\log n\right) \\
&\quad\leq 2\exp\left(-c_9 \min\left(\frac{k^2}{\log^2 n\left((k/d)^2+1\right)\|P_2\|_F^2}, \frac{k}{\log n\,(k/d+1)\,\|P_2\|_{\mathsf{op}}}\right)\right) \\
&\quad\leq 2\exp\left(-c_9 \min\left(\frac{k}{\log^2 n\left((k/d)^2+1\right)}, \frac{k}{\log n\,(k/d+1)}\right)\right) \\
&\quad\leq \exp\left(-c_{10}\log^2 n\right),
\end{aligned} \tag{D.84}$$

where the last inequality comes from Assumption 3.4. This, together with (D.79), (D.77), (D.80), and (D.83), proves the second part of the statement, and therefore the desired result.

□

Finally, we are ready to give the proof of Theorem 3.6.

Proof of Theorem 3.6. We will prove the statement for the following definition of γ_{RF}, independent from z_1 and z_1^s,

$$\gamma_{\mathsf{RF}} := \frac{\mathbb{E}_{z_1,z_1^s}\left[\varphi(z_1^s)^\top P_{\Phi_{-1}}^\perp \varphi(z_1)\right]}{\mathbb{E}_{z_1}\left[\left\|P_{\Phi_{-1}}^\perp \varphi(z_1)\right\|_2^2\right]}. \tag{D.85}$$

By Lemma D.1 and D.7, we have

$$\left\|P_{\Phi_{-1}}^\perp \varphi(z)\right\|_2^2 = \Omega(k) \tag{D.86}$$

with probability at least $1 - \exp(-c_1 \log^2 n)$ over V, Z_{-1} and z. This, together with Lemma D.12, gives

$$\left|\frac{\varphi(z_1^s)^\top P_{\Phi_{-1}}^\perp \varphi(z_1)}{\left\|P_{\Phi_{-1}}^\perp \varphi(z_1)\right\|_2^2} - \frac{\mathbb{E}_{z_1,z_1^s}\left[\varphi(z_1^s)^\top P_{\Phi_{-1}}^\perp \varphi(z_1)\right]}{\mathbb{E}_{z_1}\left[\left\|P_{\Phi_{-1}}^\perp \varphi(z_1)\right\|_2^2\right]}\right| = o(1), \tag{D.87}$$

with probability at least $1 - \exp(-c_2 \log^2 n)$ over V, Z and x, which proves the first part of the statement.

The upper-bound on γ_{RF} can be obtained applying Cauchy-Schwarz twice

$$\begin{aligned}\frac{\mathbb{E}_{z_1,z_1^s}\left[\varphi(z_1^s)^\top P^\perp_{\Phi_{-1}}\varphi(z_1)\right]}{\mathbb{E}_{z_1}\left[\left\|P^\perp_{\Phi_{-1}}\varphi(z_1)\right\|_2^2\right]} &\leq \frac{\mathbb{E}_{z_1,z_1^s}\left[\left\|P^\perp_{\Phi_{-1}}\varphi(z_1^s)\right\|_2\left\|P^\perp_{\Phi_{-1}}\varphi(z_1)\right\|_2\right]}{\mathbb{E}_{z_1}\left[\left\|P^\perp_{\Phi_{-1}}\varphi(z_1)\right\|_2^2\right]} \\ &\leq \frac{\sqrt{\mathbb{E}_{z_1^s}\left[\left\|P^\perp_{\Phi_{-1}}\varphi(z_1^s)\right\|_2^2\right]}\sqrt{\mathbb{E}_{z_1}\left[\left\|P^\perp_{\Phi_{-1}}\varphi(z_1)\right\|_2^2\right]}}{\mathbb{E}_{z_1}\left[\left\|P^\perp_{\Phi_{-1}}\varphi(z_1)\right\|_2^2\right]} = 1.\end{aligned} \tag{D.88}$$

Let's now focus on the lower bound. By Assumption3.4 and Lemma D.5 (in which we consider the degenerate case $\alpha = 1$ and set $t = n$), we have

$$\left|\left\|\tilde{\phi}\left(Vz_1\right)\right\|_2^2 - k\sum_{l=1}^{+\infty}\mu_l^2\right| = o(k), \tag{D.89}$$

with probability at least $1 - \exp(-c_3\log^2 n)$ over V and z_1. Then, a few applications of the triangle inequality give

$$\begin{aligned}\frac{\mathbb{E}_{z_1,z_1^s}\left[\varphi(z_1^s)^\top P^\perp_{\Phi_{-1}}\varphi(z_1)\right]}{\mathbb{E}_{z_1}\left[\left\|P^\perp_{\Phi_{-1}}\varphi(z_1)\right\|_2^2\right]} &\geq \frac{\varphi(z_1^s)^\top P^\perp_{\Phi_{-1}}\varphi(z_1)}{\left\|P^\perp_{\Phi_{-1}}\varphi(z_1)\right\|_2^2} - o(1) \\ &\geq \frac{\varphi(z_1^s)^\top P^\perp_{\Phi_{-1}}\varphi(z_1)}{\left\|P^\perp_{\tilde{\Phi}_{-1}}\tilde{\varphi}(z_1)\right\|_2^2} - o(1) \\ &\geq \frac{k\left(\sum_{l=2}^{+\infty}\mu_l^2\alpha^l\right) + \mu_1^2\left\|P^\perp_{\Phi_{-1}}V_y y_1\right\|_2^2}{\|\tilde{\varphi}(z_1)\|_2^2} - o(1) \\ &\geq \frac{k\left(\sum_{l=2}^{+\infty}\mu_l^2\alpha^l\right) + \mu_1^2\left\|P^\perp_{\Phi_{-1}}V_y y_1\right\|_2^2}{k\sum_{l=1}^{+\infty}\mu_l^2} - o(1) \\ &\geq \frac{\sum_{l=2}^{+\infty}\mu_l^2\alpha^l}{\sum_{l=1}^{+\infty}\mu_l^2} - o(1),\end{aligned} \tag{D.90}$$

where the first inequality is a consequence of (D.87), the second of Lemma D.8 and (D.86), the third of Lemma D.11 and again (D.86), and the fourth of (D.89), and they jointly hold with probability $1 - \exp(-c_4\log^2 n)$ over V, Z_{-1} and z_1. Again, as the statement does not depend on z_1, we can conclude that it holds with the same probability only over the probability spaces of V and Z_{-1}, and the thesis readily follows. □

D.4 Proofs for NTK features

In this section, we will indicate with $Z \in \mathbb{R}^{n\times d}$ the data matrix, such that its rows are sampled independently from $\mathcal{P}_Z$ (see Assumption 3.3). We denote by $W \in \mathbb{R}^{k\times d}$ the weight matrix at initialization, such that $W_{ij} \sim_{\text{i.i.d.}} \mathcal{N}(0, 1/d)$. Thus, the feature map is given by (see (3.24))

$$\varphi(z) := z \otimes \phi'(Wz) \in \mathbb{R}^{dk}, \tag{D.91}$$

where ϕ' is the derivative of the activation function ϕ, applied component-wise to the vector Wz. We use the shorthands $\Phi := Z * \phi'(ZW^\top) \in \mathbb{R}^{n\times p}$ and $K := \Phi\Phi^\top \in \mathbb{R}^{n\times n}$, where $*$

denotes the Khatri-Rao product, defined in Appendix B. We indicate with $\Phi_{-1} \in \mathbb{R}^{(n-1)\times k}$ the matrix Φ without the first row, and we define $K_{-1} := \Phi_{-1}\Phi_{-1}^\top$. We call P_Φ the projector over the span of the rows of Φ, and $P_{\Phi_{-1}}$ the projector over the span of the rows of Φ_{-1}. We use the notations $\tilde{\varphi}(z) := \varphi(z) - \mathbb{E}_W[\varphi(z)]$ and $\tilde{\Phi}_{-1} := \Phi_{-1} - \mathbb{E}_W[\Phi_{-1}]$ to indicate the centered feature map and matrix respectively, where the centering is with respect to W. We indicate with μ_l' the l-th Hermite coefficient of ϕ'. We use the notation $z_1^s = [x, y_1]$, where $x \sim \mathcal{P}_X$ is sampled independently from V and Z. We define $\alpha = d_y/d$. Throughout this section, for compactness, we drop the subscripts "NTK" from these quantities, as we will only treat the proofs related to Section 3.6. Again for the sake of compactness, we will not re-introduce such quantities in the statements or the proofs of the following lemmas.

The content of this section can be summarized as follows:

- In Lemma D.13, we prove the lower bound on the smallest eigenvalue of K, adapting to our settings the main result of Bombari et al. [2022b].
- In Lemma D.17, we treat separately a term that derives from $\mathbb{E}_W[\phi'(Wz)] = \mu_0'\mathbf{1}_k$, showing that we can *center* the derivative of the activation function (Lemma D.21), without changing our final statement in Theorem 3.9. This step is necessary only if $\mu_0' \neq 0$. Our proof tackles the problem proving the thesis on a set of "perturbed" inputs $\bar{Z}_{-1}(\delta)$ (Lemma D.16), critically exploiting the non degenerate behaviour of their rows (Lemma D.15), and transfers the result on the original term, using continuity arguments with respect to the perturbation (Lemma D.14).
- In Lemma D.20, we show that the centered features $\tilde{\varphi}(z_1)$ and $\tilde{\varphi}(z_1^s)$ have a negligible component in the space spanned by the rows of Φ_{-1}. To achieve this, we exploit the bound proved in Lemma D.19.
- To conclude, we prove Theorem 3.9, exploiting also the concentration result provided in Lemma D.18.

Lemma D.13. *We have that*

$$\lambda_{\min}(K) = \Omega(kd), \tag{D.92}$$

with probability at least $1 - ne^{-c\log^2 k} - e^{-c\log^2 n}$ *over* Z *and* W*, where* c *is an absolute constant.*

Proof. The result follows from Theorem 3.1 of Bombari et al. [2022b]. Notice that our assumptions on the data distribution $\mathcal{P}_Z$ are stronger, and that our initialization of the very last layer (which differs from the Gaussian initialization in Bombari et al. [2022b]) does not change the result. Assumption 3.7, *i.e.*, $k = O(d)$, satisfies the *loose pyramidal topology* condition (cf. Assumption 2.4 in Bombari et al. [2022b]), and Assumption 3.7 is the same as Assumption 2.5 in Bombari et al. [2022b]. An important difference is that we do not assume the activation function ϕ to be Lipschitz anymore. This, however, stops being a necessary assumption since we are working with a 2-layer neural network, and ϕ doesn't appear in the expression of NTK. □

Lemma D.14. *Let* $A \in \mathbb{R}^{(n-1)\times d}$ *be a generic matrix, and let* $\bar{Z}_{-1}(\delta)$ *and* $\bar{\Phi}_{-1}(\delta)$ *be defined as*

$$\bar{Z}_{-1}(\delta) := Z_{-1} + \delta A, \tag{D.93}$$

$$\bar{\Phi}_{-1}(\delta) := \bar{Z}_{-1}(\delta) * \phi'\left(Z_{-1}W^\top\right). \tag{D.94}$$

Let $\bar{P}_{\Phi_{-1}}(\delta) \in \mathbb{R}^{dk \times dk}$ be the projector over the $Span$ of the rows of $\bar{\Phi}_{-1}(\delta)$. Then, we have that $\bar{P}^{\perp}_{\Phi_{-1}}(\delta)$ is continuous in $\delta = 0$ with probability at least $1 - ne^{-c\log^2 k} - e^{-c\log^2 n}$ over Z and W, where c is an absolute constant and where the continuity is with respect to $\|\cdot\|_{\mathsf{op}}$.

Proof. In this proof, when we say that a matrix is continuous with respect to δ, we always intend with respect to the operator norm $\|\cdot\|_{\mathsf{op}}$. Then, $\bar{\Phi}_{-1}(\delta)$ is continuous in 0, as

$$\left\|\bar{\Phi}_{-1}(\delta) - \bar{\Phi}_{-1}(0)\right\|_{\mathsf{op}} = \left\|\delta A * \phi'\left(Z_{-1}W^{\top}\right)\right\|_{\mathsf{op}} \leq \delta \left\|A\right\|_{\mathsf{op}} \max_{2\leq i\leq n} \left\|\phi'\left(Wz_i\right)\right\|_2, \quad \text{(D.95)}$$

where the second step follows from Equation (3.7.13) in Johnson [1990].

By Weyl's inequality, this also implies that $\lambda_{\min}\left(\bar{\Phi}_{-1}(\delta)\bar{\Phi}_{-1}(\delta)^{\top}\right)$ is continuous in $\delta = 0$. Recall that, by Lemma D.13, $\det\left(\bar{\Phi}_{-1}(0)\bar{\Phi}_{-1}(0)^{\top}\right) \equiv \det\left(\Phi_{-1}\Phi_{-1}^{\top}\right) \neq 0$ with probability at least $1 - ne^{-c\log^2 k} - e^{-c\log^2 n}$ over Z and W. This implies that $\left(\bar{\Phi}_{-1}(\delta)\bar{\Phi}_{-1}(\delta)^{\top}\right)^{-1}$ is also continuous, as for every invertible matrix M we have $M^{-1} = \mathsf{Adj}(M)/\det(M)$ (where $\mathsf{Adj}(M)$ denotes the Adjugate of the matrix M), and both $\mathsf{Adj}(\cdot)$ and $\det(\cdot)$ are continuous mappings. Thus, as $\bar{P}_{\Phi_{-1}}(0) = \bar{\Phi}_{-1}(0)^{\top}\left(\bar{\Phi}_{-1}(0)\bar{\Phi}_{-1}(0)^{\top}\right)^{-1}\bar{\Phi}_{-1}(0)$ (see (D.2)), we also have the continuity of $\bar{P}_{\Phi_{-1}}(\delta)$ in $\delta = 0$, which gives the thesis. □

Lemma D.15. *Let $A \in \mathbb{R}^{(n-1)\times d}$ be a matrix with entries sampled independently (between each other and from everything else) from a standard Gaussian distribution. Then, for every $\delta > 0$, with probability 1 over A, the rows of $\bar{Z}_{-1} := Z_{-1} + \delta A$ span $\mathbb{R}^d$.*

Proof. As $n - 1 \geq d$, by Assumption 3.7,negating the thesis would imply that the rows of $\bar{Z}_{-1}$ are linearly dependent, and that they belong to a subspace with dimension at most $d-1$. This would imply that there exists a row of $\bar{Z}_{-1}$, call it $\bar{z}_j$, such that $\bar{z}_j$ belongs to the space spanned by all the other rows of $\bar{Z}_{-1}$, with dimension at most $d-1$. This means that $A_{j:}$ has to belong to an affine space with the same dimension, which we can consider fixed, as it's not a function of the random vector $A_{j:}$, but only of Z_{-1} and $\{A_{i:}\}_{i\neq j}$. As the entries of $A_{j:}$ are sampled independently from a standard Gaussian distribution, this happens with probability 0. □

Lemma D.16. *Let $A \in \mathbb{R}^{(n-1)\times d}$ be a matrix with entries sampled independently (between each other and from everything else) from a standard Gaussian distribution. Let $\bar{Z}_{-1}(\delta) := Z_{-1} + \delta A$ and $\bar{\Phi}_{-1}(\delta) := \bar{Z}_{-1}(\delta) * \phi'\left(Z_{-1}W^{\top}\right)$. Let $\bar{P}_{\Phi_{-1}}(\delta) \in \mathbb{R}^{dk\times dk}$ be the projector over the $Span$ of the rows of $\bar{\Phi}_{-1}(\delta)$. Let $\mu_0' \neq 0$. Then, for $z \sim \mathcal{P}_Z$, and for any $\delta > 0$, we have*

$$\left\|\bar{P}^{\perp}_{\Phi_{-1}}(\delta)\left(z \otimes \mathbf{1}_k\right)\right\|_2 = o(\sqrt{dk}), \quad \text{(D.96)}$$

with probability at least $1 - \exp(-c\log^2 n)$ over Z, W, and A, where c is an absolute constant.

Proof. Let $B_{-1} := \phi'(Z_{-1}W^{\top}) \in \mathbb{R}^{(n-1)\times k}$. Notice that, for any $\zeta \in \mathbb{R}^{n-1}$, the following identity holds

$$\bar{\Phi}_{-1}^{\top}(\delta)\zeta = \left(\bar{Z}_{-1}(\delta) * B_{-1}\right)^{\top}\zeta = \left(\bar{Z}_{-1}^{\top}(\delta)\zeta\right) \otimes \left(B_{-1}^{\top}\mathbf{1}_{n-1}\right). \quad \text{(D.97)}$$

Note that $B_{-1}^\top = \mu_0' \mathbf{1}_k \mathbf{1}_{n-1}^\top + \tilde{B}_{-1}^\top$, where $\tilde{B}_{-1}^\top = \phi'(WZ_{-1}^\top) - \mathbb{E}_W\left[\phi'(WZ_{-1}^\top)\right]$ is a $k \times (n-1)$ matrix with i.i.d. and mean-0 rows. For an argument equivalent to the one used for (D.48) and (D.49), we have

$$\left\|\tilde{B}_{-1}^\top\right\|_{\mathsf{op}} = O\left(\left(\sqrt{k}+\sqrt{n}\right)\left(\sqrt{n/d}+1\right)\right), \tag{D.98}$$

with probability at least $1 - \exp(-c\log^2 n)$ over Z_{-1} and W. Thus, we can write

$$B_{-1}^\top \frac{\mathbf{1}_{n-1}}{\mu_0'(n-1)} = \left(\mu_0' \mathbf{1}_k \mathbf{1}_{n-1}^\top + \tilde{B}_{-1}^\top\right)\frac{\mathbf{1}_{n-1}}{\mu_0'(n-1)} = \mathbf{1}_k + \tilde{B}_{-1}^\top \frac{\mathbf{1}_{n-1}}{\mu_0'(n-1)} =: \mathbf{1}_k + v, \tag{D.99}$$

where we have

$$\|v\|_2 \le \left\|\tilde{B}_{-1}^\top\right\|_{\mathsf{op}} \left\|\frac{\mathbf{1}_{n-1}}{\mu_0'(n-1)}\right\|_2 = O\left(\left(\sqrt{k/n}+1\right)\left(\sqrt{n/d}+1\right)\right) = o(\sqrt{k}), \tag{D.100}$$

where the last step is a consequence of Assumption 3.7. Plugging (D.99) in (D.97) we get

$$\frac{1}{\mu_0'(n-1)}\bar{\Phi}_{-1}^\top(\delta)\zeta = \frac{1}{\mu_0'(n-1)}\left(\bar{Z}_{-1}(\delta) * B_{-1}\right)^\top \zeta = \left(\bar{Z}_{-1}^\top(\delta)\zeta\right) \otimes (\mathbf{1}_k + v). \tag{D.101}$$

By Lemma D.15, we have that the rows of $\bar{Z}_{-1}(\delta)$ span $\mathbb{R}^d$, with probability 1 over A. Thus, conditioning on this event, we can set ζ to be a vector such that $z = \bar{Z}_{-1}^\top(\delta)\zeta$. We can therefore rewrite the previous equation as

$$\frac{1}{\mu_0'(n-1)}\bar{\Phi}_{-1}^\top(\delta)\zeta = z \otimes \mathbf{1}_k + z \otimes v. \tag{D.102}$$

Thus, we can conclude

$$\begin{aligned}
\left\|\bar{P}_{\Phi_{-1}}^\perp(\delta)\,(z \otimes \mathbf{1}_k)\right\|_2 &= \left\|P_{\Phi_{-1}}^\perp\left(\frac{\Phi_{-1}^\top(\delta)\zeta}{\mu_0'(n-1)} - z \otimes v\right)\right\|_2 \\
&\le \left\|\bar{P}_{\Phi_{-1}}^\perp(\delta)\Phi_{-1}^\top(\delta)\frac{\zeta}{\mu_0'(n-1)}\right\|_2 + \|z \otimes v\|_2 \\
&= \|z\|_2 \|v\|_2 = o(\sqrt{dk}),
\end{aligned} \tag{D.103}$$

where in the second step we use the triangle inequality, in the third step we use that $\Phi_{-1}^\top(\delta) = \bar{P}_{\Phi_{-1}}(\delta)\Phi_{-1}^\top(\delta)$, and in the last step we use (D.100). The desired result readily follows. □

Lemma D.17. *Let $\mu_0' \neq 0$. Then, for any $z \in \mathbb{R}^d$, we have*

$$\left\|P_{\Phi_{-1}}^\perp\,(z \otimes \mathbf{1}_k)\right\|_2 = o(\sqrt{dk}), \tag{D.104}$$

with probability at least $1 - ne^{-c\log^2 k} - e^{-c\log^2 n}$ over Z and W, where c is an absolute constant.

Proof. Let $A \in \mathbb{R}^{(n-1)\times d}$ be a matrix with entries sampled independently (between each other and from everything else) from a standard Gaussian distribution. Let $\bar{Z}_{-1}(\delta) := Z_{-1} + \delta A$

and $\bar{\Phi}_{-1}(\delta) := \bar{Z}_{-1}(\delta) * \phi'\left(Z_{-1}W^\top\right)$. Let $\bar{P}_{\Phi_{-1}}(\delta) \in \mathbb{R}^{dk\times dk}$ be the projector over the Span of the rows of $\bar{\Phi}_{-1}(\delta)$.

By triangle inequality, we can write

$$\left\|P^\perp_{\Phi_{-1}}\left(z\otimes \mathbf{1}_k\right)\right\|_2 \le \left\|P^\perp_{\Phi_{-1}} - \bar{P}^\perp_{\Phi_{-1}}(\delta)\right\|_{\mathsf{op}} \|z\otimes \mathbf{1}_k\|_2 + \left\|\bar{P}^\perp_{\Phi_{-1}}(\delta)\left(z\otimes \mathbf{1}_k\right)\right\|_2. \tag{D.105}$$

Because of Lemma D.14, with probability at least $1 - ne^{-c\log^2 k} - e^{-c\log^2 n}$ over Z and W, $\bar{P}^\perp_{\Phi_{-1}}(\delta)$ is continuous in $\delta = 0$, with respect to $\|\cdot\|_{\mathsf{op}}$. Thus, there exists $\delta^* > 0$ such that, for every $\delta \in [0, \delta^*]$,

$$\left\|P^\perp_{\Phi_{-1}} - \bar{P}^\perp_{\Phi_{-1}}(\delta)\right\|_{\mathsf{op}} \equiv \left\|\bar{P}^\perp_{\Phi_{-1}}(0) - \bar{P}^\perp_{\Phi_{-1}}(\delta)\right\|_{\mathsf{op}} < \frac{1}{n}. \tag{D.106}$$

Hence, setting $\delta = \delta^*$ in (D.105), we get

$$\begin{aligned}\left\|P^\perp_{\Phi_{-1}}\left(z\otimes \mathbf{1}_k\right)\right\|_2 &\le \left\|P^\perp_{\Phi_{-1}} - \bar{P}^\perp_{\Phi_{-1}}(\delta^*)\right\|_{\mathsf{op}} \|z\otimes \mathbf{1}_k\|_2 + \left\|\bar{P}^\perp_{\Phi_{-1}}(\delta^*)\left(z\otimes \mathbf{1}_k\right)\right\|_2 \\ &\le \|z\|_2 \|\mathbf{1}_k\|_2 / n + \left\|\bar{P}^\perp_{\Phi_{-1}}(\delta^*)\left(z\otimes \mathbf{1}_k\right)\right\|_2 \\ &= o(\sqrt{dk}),\end{aligned} \tag{D.107}$$

where the last step is a consequence of Lemma D.16, and it holds with probability at least $1 - \exp(-c\log^2 n)$ over Z, W, and A. As the LHS of the previous equation doesn't depend on A, the statements holds with the same probability, just over the probability spaces of Z and W, which gives the desired result. □

Lemma D.18. *We have*

$$\left|\frac{\tilde{\varphi}(z_1^s)^\top \tilde{\varphi}(z_1)}{\|\tilde{\varphi}(z_1)\|_2^2} - \alpha\,\frac{\sum_{l=1}^{+\infty}\mu_l'^2\alpha^i}{\sum_{l=1}^{+\infty}\mu_l'^2}\right| = o(1), \tag{D.108}$$

with probability at least $1 - \exp(-c\log^2 n) - \exp(-c\log^2 k)$ *over* W *and* z_1*, where* c *is an absolute constant. With the same probability, we also have*

$$\tilde{\varphi}(z_1^s)^\top \tilde{\varphi}(z_1) = \Theta(dk), \qquad \|\tilde{\varphi}(z_1)\|_2^2 = \Theta(dk). \tag{D.109}$$

Proof. We have

$$\|\tilde{\varphi}(z_1)\|_2^2 = \left\|z_1 \otimes \tilde{\phi}'\left(Wz_1\right)\right\|_2^2 = \|z_1\|_2^2 \left\|\tilde{\phi}'\left(Wz_1\right)\right\|_2^2 = d\left\|\tilde{\phi}'\left(Wz_1\right)\right\|_2^2. \tag{D.110}$$

By Assumption 3.7 and Lemma D.5 (in which we consider the degenerate case $\alpha = 1$ and set $t = k$), we have

$$\left|\left\|\tilde{\phi}'\left(Wz_1\right)\right\|_2^2 - k\sum_{l=1}^{+\infty}\mu_l'^2\right| = o(k), \tag{D.111}$$

with probability at least $1 - \exp(-c\log^2 k)$ over W and z_1. Thus, we have

$$\left|\|\tilde{\varphi}(z_1)\|_2^2 - dk\sum_{l=1}^{+\infty}\mu_l'^2\right| = o(dk). \tag{D.112}$$

Notice that the second term in the modulus is $\Theta(dk)$, since the μ_l'-s cannot be all 0, because of Assumption 3.8; this shows that $\|\tilde{\varphi}(z_1)\|_2^2 = \Theta(dk)$.

Similarly, we can write

$$\tilde{\varphi}(z_1^s)^\top \tilde{\varphi}(z_1) = \left(z_1^\top z_1^s\right) \left(\tilde{\phi}'\left(Wz_1\right)^\top \tilde{\phi}'\left(Wz_1^s\right)\right). \tag{D.113}$$

We have

$$\left|z_1^\top z_1^s - \alpha d\right| = \left|x_1^\top x\right| \leq \sqrt{d_x} \log d = o(d), \tag{D.114}$$

where the inequality holds with probability at least $1 - \exp(-c_1 \log^2 d) \geq 1 - \exp(-c_2 \log^2 n)$ over x_1, as we are taking the inner product of two independent and sub-Gaussian vectors with norm $\sqrt{d_x}$. Furthermore, again by Assumption3.7 and Lemma D.5, we have

$$\left|\tilde{\phi}'\left(Wz_1\right)^\top \tilde{\phi}'\left(Wz_1^s\right) - k \sum_{l=1}^{+\infty} {\mu_l'}^2 \alpha^l\right| = o(k), \tag{D.115}$$

with probability at least $1 - \exp(-c_3 \log^2 k)$ over W and z_1. Notice that the second term in the modulus is $\Theta(k)$, because of Assumption 3.8.

Thus, putting (D.113), (D.114) and (D.115) together, we get

$$\left|\tilde{\varphi}(z_1^s)^\top \tilde{\varphi}(z_1) - dk\alpha \sum_{l=1}^{+\infty} {\mu_l'}^2 \alpha^l\right| = o(dk), \tag{D.116}$$

with probability at least $1 - \exp(-c_3 \log^2 k) - \exp(-c_2 \log^2 n)$ over W and z_1; this shows that $\tilde{\varphi}(z_1^s)^\top \tilde{\varphi}(z_1) = \Theta(dk)$.

Finally, merging (D.116) with (D.112) and applying triangle inequality, (D.108) follows and the proof is complete. □

Lemma D.19. *Let $z \sim \mathcal{P}_Z$ be sampled independently from Z_{-1}. Then,*

$$\left\|\Phi_{-1}\tilde{\varphi}(z)\right\|_2 = o\left(dk\right), \tag{D.117}$$

with probability at least $1 - \exp(-c \log^2 n)$ over W and z, where c is an absolute constant.

Proof. Let's look at the i-th entry of the vector $\Phi_{-1}\tilde{\varphi}(z)$, *i.e.*,

$$\varphi(z_{i+1})^\top \tilde{\varphi}(z) = \left(z_{i+1}^\top z\right) \left(\phi'(Wz_{i+1})^\top \tilde{\phi}'(Wz)\right). \tag{D.118}$$

As z and z_{i+1} are sub-Gaussian and independent with norm $\sqrt{d}$, we can write $\left|z^\top z_{i+1}\right| = O\left(\sqrt{d} \log n\right)$ with probability at least $1 - \exp(-c \log^2 n)$ over z. We will condition on such high probability event until the end of the proof.

By Lemma D.4, setting $t = n$, we have

$$\left|\phi'(Wz_{i+1})^\top \tilde{\phi}'(Wz) - \mathbb{E}_W\left[\phi'(Wz_{i+1})^\top \tilde{\phi}'(Wz)\right]\right| = O\left(\sqrt{k} \log n\right), \tag{D.119}$$

with probability at least $1-\exp(-c_1\log^2 n)$ over W. Exploiting the Hermite expansion of ϕ' and $\tilde{\phi}'$, we have

$$\begin{aligned}\left|\mathbb{E}_W\left[\phi'(Wz_{i+1})^\top\tilde{\phi}'(Wz)\right]\right| &\le k\sum_{l=1}^{+\infty}\mu_l'^2\left(\frac{\left|z_{i+1}^\top z\right|}{\|z_{i+1}\|_2\|z\|_2}\right)^l\\ &\le k\max_l\mu_l'^2\sum_{l=1}^{+\infty}\left(\frac{\left|z_{i+1}^\top z\right|}{\|z_{i+1}\|_2\|z\|_2}\right)^l\\ &= k\max_l\mu_l'^2\frac{\left|z_{i+1}^\top z\right|}{\|z_{i+1}\|_2\|z\|_2}\frac{1}{1-\frac{\left|z_{i+1}^\top z\right|}{\|z_{i+1}\|_2\|z\|_2}}\\ &\le 2k\max_l\mu_l'^2\frac{\left|z_{i+1}^\top z\right|}{\|z_{i+1}\|_2\|z\|_2} = O\left(\frac{k\log n}{\sqrt{d}}\right).\end{aligned} \tag{D.120}$$

Putting together (D.119) and (D.120), and applying triangle inequality, we get

$$\left|\phi'(Wz_{i+1})^\top\tilde{\phi}'(Wz)\right| = O\left(\sqrt{k}\log n+\frac{k\log n}{\sqrt{d}}\right) = O\left(\sqrt{k}\log n\right), \tag{D.121}$$

where the last step is a consequence of Assumption 3.7.Comparing this last result with (D.118), we obtain

$$\left|\varphi(z_{i+1})^\top\tilde{\varphi}(z)\right| = O\left(\sqrt{dk}\log^2 n\right), \tag{D.122}$$

with probability at least $1-\exp(-c_2\log^2 n)$ over W and z.

We want the previous equation to hold for all $1\le i\le n-1$. Performing a union bound, we have that this is true with probability at least $1-(n-1)\exp(-c_2\log^2 n)\ge 1-\exp(-c_3\log^2 n)$ over W and z. Thus, with such probability, we have

$$\begin{aligned}\|\Phi_{-1}\tilde{\varphi}(z)\|_2 &\le \sqrt{n-1}\max_i\left|\varphi(z_{i+1})^\top\tilde{\varphi}(z)\right|\\ &= O\left(\sqrt{dk}\sqrt{n}\log^2 n\right) = o(dk),\end{aligned} \tag{D.123}$$

where the last step follows from Assumption 3.7. □

Lemma D.20. *We have*

$$\left|\frac{\tilde{\varphi}(z_1^s)^\top\tilde{\varphi}(z_1)-\tilde{\varphi}(z_1^s)^\top P_{\Phi_{-1}}\tilde{\varphi}(z_1)}{\left\|\tilde{\varphi}(z_1)-P_{\Phi_{-1}}\tilde{\varphi}(z_1)\right\|_2^2}-\frac{\tilde{\varphi}(z_1^s)^\top\tilde{\varphi}(z_1)}{\|\tilde{\varphi}(z_1)\|_2^2}\right| = o(1), \tag{D.124}$$

with probability at least $1-n\exp(-c\log^2 k)-\exp(-c\log^2 n)$ *over* Z, x *and* W, *where* c *is an absolute constant. With the same probability, we also have*

$$\tilde{\varphi}(z_1^s)^\top\tilde{\varphi}(z_1)-\tilde{\varphi}(z_1^s)^\top P_{\Phi_{-1}}\tilde{\varphi}(z_1) = \Theta(dk), \qquad \left\|\tilde{\varphi}(z_1)-P_{\Phi_{-1}}\tilde{\varphi}(z_1)\right\|_2^2 = \Theta(dk). \tag{D.125}$$

Proof. Notice that, with probability at least $1-\exp(-c\log^2 n)-\exp(-c\log^2 k)$ over W and z_1, we have both

$$\tilde{\varphi}(z_1^s)^\top\tilde{\varphi}(z_1) = \Theta(dk) \qquad \|\tilde{\varphi}(z_1)\|_2^2 = \Theta(dk). \tag{D.126}$$

by the second statement of Lemma D.18. Furthermore,

$$\begin{aligned}\left|\tilde{\varphi}(z_1^s)^\top P_{\Phi_{-1}}\tilde{\varphi}(z_1)\right| &= \left|\tilde{\varphi}(z_1^s)^\top \Phi_{-1}^\top K_{-1}^{-1}\Phi_{-1}\tilde{\varphi}(z_1)\right| \\ &\leq \|\Phi_{-1}\tilde{\varphi}(z_1^s)\|_2 \,\lambda_{\min}\left(K_{-1}\right)^{-1}\|\Phi_{-1}\tilde{\varphi}(z_1)\|_2 \\ &= o(dk)O\left(\frac{1}{dk}\right)o(dk) = o(dk),\end{aligned} \tag{D.127}$$

where the third step is justified by Lemmas D.13 and D.19, and holds with probability at least $1 - ne^{-c\log^2 k} - e^{-c\log^2 n}$ over Z, x, and W. A similar argument can be used to show that $\left\|P_{\Phi_{-1}}\tilde{\varphi}(z_1)\right\|_2^2 = o(dk)$, which, together with (D.127) and (D.126), and a straightforward application of the triangle inequality, provides the thesis. □

Lemma D.21. *We have*

$$\left|\frac{\varphi(z_1^s)^\top P_{\Phi_{-1}}^\perp \varphi(z_1)}{\left\|P_{\Phi_{-1}}^\perp \varphi(z_1)\right\|_2^2} - \frac{\tilde{\varphi}(z_1^s)^\top P_{\Phi_{-1}}^\perp \tilde{\varphi}(z_1)}{\left\|P_{\Phi_{-1}}^\perp \tilde{\varphi}(z_1)\right\|_2^2}\right| = o(1), \tag{D.128}$$

with probability at least $1 - n\exp(-c\log^2 k) - \exp(-c\log^2 n)$ *over* Z, x *and* W, *where* c *is an absolute constant.*

Proof. If $\mu_0' = 0$, the thesis is trivial, as $\varphi \equiv \tilde{\varphi}$. If $\mu_0' \neq 0$, we can apply Lemma D.17, and the proof proceeds as follows.

First, we notice that the second term in the modulus in the statement corresponds to the first term in the statement of Lemma D.20. We will condition on the result of Lemma D.20 to hold until the end of the proof. Notice that this also implies

$$\tilde{\varphi}(z_1^s)^\top P_{\Phi_{-1}}^\perp \tilde{\varphi}(z_1) = \Theta(dk), \qquad \left\|P_{\Phi_{-1}}^\perp \tilde{\varphi}(z_1)\right\|_2^2 = \Theta(dk), \tag{D.129}$$

with probability at least $1 - n\exp(-c\log^2 k) - \exp(-c\log^2 n)$ over Z, x, and W. Due to Lemma D.17, we jointly have

$$\left\|P_{\Phi_{-1}}^\perp (z_1 \otimes \mathbf{1}_k)\right\|_2 = o(\sqrt{dk}), \qquad \left\|P_{\Phi_{-1}}^\perp (z_1^s \otimes \mathbf{1}_k)\right\|_2 = o(\sqrt{dk}), \tag{D.130}$$

with probability at least $1 - \exp(c\log^2 n)$ over Z_{-1} and W. Also, by Lemma D.4 and Assumption 3.4, we jointly have

$$\left\|P_{\Phi_{-1}}^\perp \varphi(z_1^s)\right\|_2 \leq \|\varphi(z_1^s)\|_2 = \|z_1^s\|_2 \,\|\phi'(W z_1^s)\|_2 = O\left(\sqrt{dk}\right), \tag{D.131}$$

and

$$\left\|P_{\Phi_{-1}}^\perp \tilde{\varphi}(z_1)\right\|_2 \leq \|\tilde{\varphi}(z_1)\|_2 = \|z_1\|_2 \left\|\tilde{\phi}'(W z_1)\right\|_2 = O\left(\sqrt{dk}\right), \tag{D.132}$$

with probability at least $1 - \exp(-c_1\log^2 n)$ over W. We will condition also on such high probability events ((D.130), (D.131), (D.132)) until the end of the proof. Thus, we can write

$$\begin{aligned}&\left|\varphi(z_1^s)^\top P_{\Phi_{-1}}^\perp \varphi(z_1) - \tilde{\varphi}(z_1^s)^\top P_{\Phi_{-1}}^\perp \tilde{\varphi}(z_1)\right| \\ &\leq \left|\varphi(z_1^s)^\top P_{\Phi_{-1}}^\perp \left(\varphi(z_1) - \tilde{\varphi}(z_1)\right)\right| + \left|\left(\varphi(z_1^s) - \tilde{\varphi}(z_1^s)\right)^\top P_{\Phi_{-1}}^\perp \tilde{\varphi}(z_1)\right| \\ &\leq \left\|P_{\Phi_{-1}}^\perp \varphi(z_1^s)\right\|_2 \left\|P_{\Phi_{-1}}^\perp (z_1 \otimes \mu_0 \mathbf{1}_k)\right\|_2 + \left\|P_{\Phi_{-1}}^\perp \tilde{\varphi}(z_1)\right\|_2 \left\|P_{\Phi_{-1}}^\perp (z_1^s \otimes \mu_0 \mathbf{1}_k)\right\|_2 = o(dk),\end{aligned} \tag{D.133}$$

where in the last step we use (D.130), (D.131), and (D.132). Similarly, we can show that

$$\begin{aligned}\left|\left\|P^{\perp}_{\Phi_{-1}}\varphi(z_1)\right\|_2 - \left\|P^{\perp}_{\Phi_{-1}}\tilde{\varphi}(z_1)\right\|_2\right| &\leq \left\|P^{\perp}_{\Phi_{-1}}\varphi(z_1) - P^{\perp}_{\Phi_{-1}}\tilde{\varphi}(z_1)\right\|_2 \\ &\leq \left\|P^{\perp}_{\Phi_{-1}}\left(z_1 \otimes \mu_0 \mathbf{1}_k\right)\right\|_2 = o(\sqrt{dk}).\end{aligned} \tag{D.134}$$

By combining (D.129), (D.133), and (D.134), the desired result readily follows. □

Finally, we are ready to give the proof of Theorem 3.9.

Proof of Theorem 3.9. We have

$$\begin{aligned}\left|\frac{\varphi(z_1^s)^\top P^{\perp}_{\Phi_{-1}}\varphi(z_1)}{\left\|P^{\perp}_{\Phi_{-1}}\varphi(z_1)\right\|_2^2} - \alpha\,\frac{\sum_{l=1}^{+\infty}\mu_l'^2\alpha^i}{\sum_{l=1}^{+\infty}\mu_l'^2}\right| &\leq \left|\frac{\varphi(z_1^s)^\top P^{\perp}_{\Phi_{-1}}\varphi(z_1)}{\left\|P^{\perp}_{\Phi_{-1}}\varphi(z_1)\right\|_2^2} - \frac{\tilde{\varphi}(z_1^s)^\top P^{\perp}_{\Phi_{-1}}\tilde{\varphi}(z_1)}{\left\|P^{\perp}_{\Phi_{-1}}\tilde{\varphi}(z_1)\right\|_2^2}\right| \\ &+ \left|\frac{\tilde{\varphi}(z_1^s)^\top\tilde{\varphi}(z_1) - \tilde{\varphi}(z_1^s)^\top P_{\Phi_{-1}}\tilde{\varphi}(z_1)}{\left\|\tilde{\varphi}(z_1) - P_{\Phi_{-1}}\tilde{\varphi}(z_1)\right\|_2^2} - \frac{\tilde{\varphi}(z_1^s)^\top\tilde{\varphi}(z_1)}{\|\tilde{\varphi}(z_1)\|_2^2}\right| \\ &+ \left|\frac{\tilde{\varphi}(z_1^s)^\top\tilde{\varphi}(z_1)}{\|\tilde{\varphi}(z_1)\|_2^2} - \alpha\,\frac{\sum_{l=1}^{+\infty}\mu_l'^2\alpha^i}{\sum_{l=1}^{+\infty}\mu_l'^2}\right| \\ &= o(1),\end{aligned} \tag{D.135}$$

where the first step is justified by the triangle inequality, and the second by Lemmas D.21, D.20, and D.18, and it holds with probability at least $1 - n\exp(-c\log^2 k) - \exp(-c\log^2 n)$ over Z, x, and W. □

D.5 Additional experiments

Figure D.1 reports the experiments on NTK features for the same setting considered in Figure 3.2 for random features. We consider binary classification tasks involving synthetic (first two plots) and standard (last two plots) datasets. As predicted by Theorem 3.9, when the number of samples n increases, the test accuracy increases and, correspondingly, the spurious accuracy decreases. Furthermore, for the synthetic dataset, while the test accuracy does not depend on α and on the activation function, the spurious accuracy increases with α and by taking an activation function with dominant low-order Hermite coefficients.

In Figure D.2, we plot the test and the spurious accuracies as a function of $0 < \alpha < 1$. While the test accuracy does not depend on α, the spurious accuracy monotonically grows with α. This is in agreement with the results of Theorems 3.6 and 3.9.

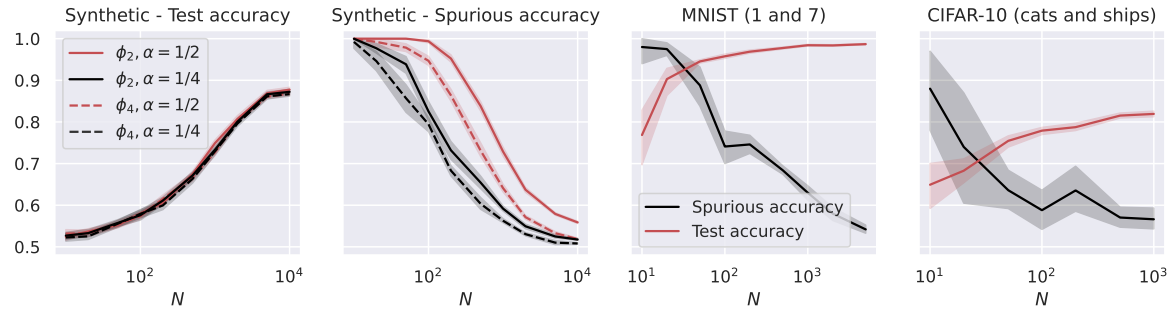


Figure D.1: Test and spurious accuracies as a function of the number of training samples n, for various binary classification tasks. In the first two plots, we consider the NTK model in (3.24) with $k = 100$ trained over Gaussian data with $d = 1000$. The labeling function is $g(x) = \text{sign}(u^\top x)$. We repeat the experiments for $\alpha = \{0.25, 0.5\}$, and for the two activations whose derivatives are $\phi_2' = h_0 + h_1$ and $\phi_4' = h_0 + h_3$, where h_i denotes the i-th Hermite polynomial (see Appendix B). In the last two plots, we consider the same model with ReLU activation, trained over two MNIST and CIFAR-10 classes. The width of the noise background is 10 pixels for MNIST and 8 pixels for CIFAR-10, see Figure 3.1. The spurious accuracy is obtained by querying the model only with the noise background from the training set, replacing all the other pixels with 0, and taking the sign of the output. As we consider binary classification, an accuracy of 0.5 is achieved by random guessing. We plot the average over 10 independent trials and the confidence band at 1 standard deviation.

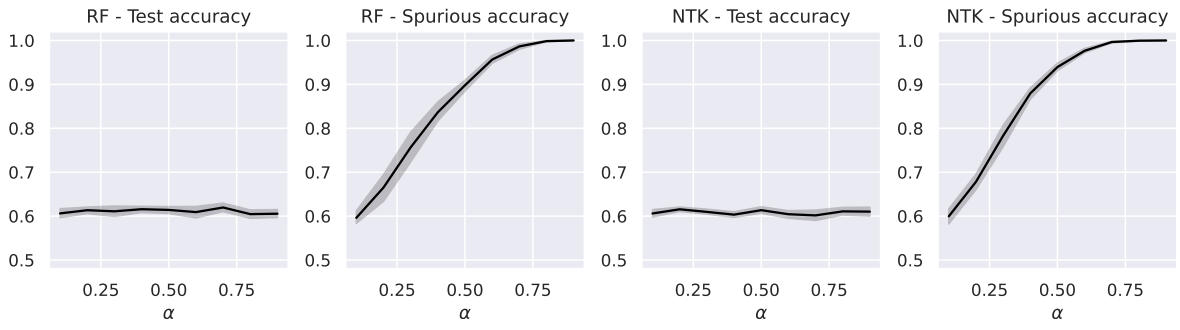


Figure D.2: Test and spurious accuracies as a function of α. We consider RF (first and second plot) and NTK (third and fourth plot) models trained on a synthetic dataset. The settings are the same as in Figures 3.2 and D.1, and we use a ReLU activation function. The number of training samples is fixed to $n = 200$.

APPENDIX E

Appendix for Chapter 4

E.1 Proof of Theorem 4.4

First, notice that, with overwhelming probability over V and X, due to Lemma 4.5 in [Bombari and Mondelli, 2025a], we have

$$\lambda_{\min}\left(\Phi\Phi^\top\right) = \Omega(p). \tag{E.1}$$

In particular, notice that the lower bound on n in their statement is not required to lower bound the smallest eigenvalue of the kernel. This implies that $\Phi\Phi^\top$ is full rank. Since for every $i \in [n]$ we have $\varphi(x_i) \in \mathrm{Span}\{\mathrm{rows}(\hat{\Phi})\}$, we also have

$$\mathrm{Span}\{\mathrm{rows}(\Phi)\} \subseteq \mathrm{Span}\{\mathrm{rows}(\hat{\Phi})\}. \tag{E.2}$$

Thus, the equation above implies that a subspace of dimension n ($\Phi\Phi^\top$ is full rank) is a subset of another subspace of dimension at most n. This implies that the two subspaces must be identical, which in turn implies, for every $i \in [n]$,

$$\varphi(\hat{x}_i) \in \mathrm{Span}\{\mathrm{rows}(\Phi)\}. \tag{E.3}$$

Throughout the proof, we omit the subscript $\hat{\imath}$ to simplify the notation and we let $\hat{x}$ denote a row of $\hat{X}$ such that $\mathcal{L}(\hat{X}) = 0$. Then, due to (E.3), we will write

$$\varphi(\hat{x}) = \Phi^\top a = \sum_{i=1}^{n} \varphi(x_i) a_i, \tag{E.4}$$

where we introduced the vector $a \in \mathbb{R}^n$ containing the coefficients of $\varphi(\hat{x})$ written in the basis $\{\varphi(x_i)\}_{i=1}^n$.

Furthermore, in this section, we will introduce the shorthand $\tilde{\varphi}(x) = \tilde{\phi}(Vx)$, where $\tilde{\phi} : \mathbb{R} \to \mathbb{R}$ is a non-linearity that shares the same Hermite coefficients as ϕ, but with the first $\tilde{\mu}_1 = 0$. This implies that $\tilde{\phi}(z) = \phi(z) - \mu_1 z$ is a Lipschitz function. We will use the shorthand $\tilde{\mu}^2 = \sum_{l=3}^{\infty} \mu_l^2$.

Lemma E.1. *For any $0 < t < p$ and any $i \in [n]$, we have that*

$$\left|\tilde{\varphi}(x_i)^\top \varphi(x_i) - \tilde{\mu}^2 p\right| = O\left(t\sqrt{p}\right), \tag{E.5}$$

with probability at least $1 - 2\exp(-ct^2)$ *over* V*. Furthermore, for any* $i \neq j$*, we have*

$$\left|\tilde{\varphi}(x_i)^\top \varphi(x_j)\right| = O\left(t\sqrt{p} + p\frac{\log^3 d}{d^{3/2}}\right), \tag{E.6}$$

with probability at least $1 - 2\exp(-ct^2) - 2\exp(-c\log^2 d)$ *over* V *and* X*. Finally, we jointly have*

$$\sup_{\hat{x}\in\sqrt{d}\mathbb{S}^{d-1}} \left|\tilde{\varphi}(\hat{x})^\top \varphi(\hat{x}) - \tilde{\mu}^2 p\right| = O\left(\sqrt{pd}\log d + \frac{p}{d^2}\right), \tag{E.7}$$

$$\sup_{\hat{x}\in\sqrt{d}\mathbb{S}^{d-1}} \left|\tilde{\varphi}(x_i)^\top \varphi(\hat{x}) - \tilde{\varphi}(x_i)^\top \tilde{\varphi}(\hat{x})\right| = O\left(\sqrt{pd}\log d + \frac{p}{d^2}\right), \tag{E.8}$$

$$\sup_{\hat{x}\in\sqrt{d}\mathbb{S}^{d-1}} \left|\varphi(x_i)^\top \tilde{\varphi}(\hat{x}) - \tilde{\varphi}(x_i)^\top \tilde{\varphi}(\hat{x})\right| = O\left(\sqrt{pd}\log d + \frac{p}{d^2}\right), \tag{E.9}$$

for all $i \in [n]$*, with probability at least* $1 - 2\exp(-cd\log^2 d)$ *over* V*.*

Proof. For the first statement, we have

$$\begin{aligned}\tilde{\varphi}(x_i)^\top \varphi(x_j) &= \sum_{k=1}^{p} \tilde{\phi}(v_k^\top x_i)\phi(v_k^\top x_j) \\ &= \sum_{k=1}^{p} \left(\tilde{\phi}(v_k^\top x_i)\phi(v_k^\top x_j) - \mathbb{E}_{v_k}\left[\tilde{\phi}(v_k^\top x_i)\phi(v_k^\top x_j)\right]\right) \\ &\quad + p\,\mathbb{E}_v\left[\tilde{\phi}(v^\top x_i)\phi(v^\top x_j)\right].\end{aligned} \tag{E.10}$$

Let us analyze the two terms in the RHS separately. The first is the sum of p independent, mean-0, sub-exponential random variables (in the probability space of V). This holds since both ϕ and $\tilde{\phi}$ are Lipschitz due to Assumption 4.2, and they arguments are sub-Gaussian. Then, due to Bernstein inequality (see Theorem 2.8.1 of [Vershynin, 2018]), for any $0 < t < p$, we have

$$\left|\sum_{k=1}^{p} \tilde{\phi}(v_k^\top x_i)\phi(v_k^\top x_j) - \mathbb{E}_{v_k}\left[\tilde{\phi}(v_k^\top x_i)\phi(v_k^\top x_j)\right]\right| \leq t\sqrt{p}, \tag{E.11}$$

with probability at least $1 - 2\exp\left(-c_1 t^2\right)$ over V. For the second term in the RHS of (E.10), using the Hermite decomposition of ϕ and $\tilde{\phi}$, we have that

$$p\,\mathbb{E}_v\left[\tilde{\phi}(v^\top x_i)\phi(v^\top x_j)\right] = p\sum_{l=3}^{+\infty} \mu_l^2 \frac{\left(x_i^\top x_j\right)^l}{d^l}. \tag{E.12}$$

Considering the case $i = j$, we readily obtain (E.5). For the case $i \neq j$, since the x_i-s are sampled independently from a sub-Gaussian distribution due to Assumption 4.1, and $\|x_i\|_2 = \sqrt{d}$ for every i, we have that

$$\max_{i\neq j} |x_i^\top x_j| \leq \sqrt{d}\log d, \tag{E.13}$$

with probability at least $1 - 2n^2\exp(-c_2\log^2 d) \geq 1 - 2\exp(-c_3\log^2 d)$, due to Assumption 4.3. Then, with this probability, we have that

$$p\sum_{l=3}^{+\infty} \mu_l^2 \frac{\left(x_i^\top x_j\right)^l}{d^l} \leq p\frac{\left(x_i^\top x_j\right)^3}{d^3}\sum_{l=3}^{+\infty}\mu_l^2 \leq \tilde{\mu}^2 p\frac{\log^3 d}{d^{3/2}}, \tag{E.14}$$

for every $i \neq j$, which gives (E.6).

Let us now consider an $\epsilon\sqrt{d}$-net of $\sqrt{d}\,\mathbb{S}^{d-1}$, namely $\{x_m^\epsilon\}_{m=1}^M$, such that for any $x \in \sqrt{d}\,\mathbb{S}^{d-1}$ there exists $m \in [M]$ such that $\|x - x_m^\epsilon\|_2 \leq \epsilon\sqrt{d}$. Due to Corollary 4.2.13 in [Vershynin, 2018], for $\epsilon < 1$ we have that the net can be chosen such that $M \leq (3/\epsilon)^d$. Notice that, for any $m \in [M]$, the same argument leading to (E.5) yields

$$\left|\tilde{\varphi}(x_m^\epsilon)^\top \varphi(x_m^\epsilon) - \tilde{\mu}^2 p\right| = O(\sqrt{pd}\log d), \tag{E.15}$$

with probability at least $1 - 2\exp(-c_3 d\log^2 d)$, after setting $t = \sqrt{d}\log d$. Setting $\epsilon = 1/d^2$, and performing a union bound on all $m \in [M]$, we have that the previous statement holds uniformly with probability at least

$$1 - 2(3/\epsilon)^d \exp(-c_3 d\log^2 d) = 1 - 2\exp((\log(3/\epsilon) - c_3\log^2 d)d) \geq 1 - 2\exp(-c_4 d\log^2 d), \tag{E.16}$$

where c_4 is an absolute constant. By Lemma C.3 in [Bombari and Mondelli, 2024a] we also have that for any $m \in [M]$, we jointly have

$$\|\varphi(x_m^\epsilon)\|_2 = O(\sqrt{p}), \qquad \|\tilde{\varphi}(x_m^\epsilon)\|_2 = O(\sqrt{p}), \tag{E.17}$$

with probability at least $1 - 2\exp(-c_5 p)$. Taking a union bound for all $m \in [M]$, (E.17) holds uniformly over the net with probability at least $1 - 2\exp(-c_6 p)$ due to Assumption 4.3. Since $\epsilon = 1/d^2$, for any $\hat{x}$, there exists m such that $\|\hat{x} - x_m^\epsilon\|_2 \leq 1/d^{3/2}$. This implies

$$\begin{aligned}
\sup_{\hat{x}\in\sqrt{d}\,\mathbb{S}^{d-1}} \|\varphi(\hat{x})\|_2 &\leq \max_{m\in[M]} \|\varphi(x_m^\epsilon)\|_2 + \sup_{\hat{x}\in\sqrt{d}\,\mathbb{S}^{d-1}} \|\varphi(\hat{x}) - \varphi(x_m^\epsilon)\|_2 \\
&\leq \max_{m\in[M]} \|\varphi(x_m^\epsilon)\|_2 + L\,\|V\|_{\mathsf{op}} \sup_{\hat{x}\in\sqrt{d}\,\mathbb{S}^{d-1}} \|\hat{x} - x_m^\epsilon\|_2 \\
&= O\left(\sqrt{p} + L\sqrt{\frac{p}{d}}\frac{1}{d^{3/2}}\right) = O(\sqrt{p}),
\end{aligned} \tag{E.18}$$

where the third step follows from $\|V\|_{\mathsf{op}} = O(\sqrt{p/d})$, which holds with probability at least $1 - 2\exp(-c_7 p)$ due to Theorem 4.4.5 in [Vershynin, 2018]. Then, we have

$$\begin{aligned}
\sup_{\hat{x}\in\sqrt{d}\,\mathbb{S}^{d-1}} \left|\tilde{\varphi}(x_m^\epsilon)^\top \varphi(x_m^\epsilon) - \tilde{\varphi}(\hat{x})^\top \varphi(\hat{x})\right| &\leq \sup_{\hat{x}\in\sqrt{d}\,\mathbb{S}^{d-1}} \left|\tilde{\varphi}(x_m^\epsilon)^\top \varphi(x_m^\epsilon) - \tilde{\varphi}(x_m^\epsilon)^\top \varphi(\hat{x})\right| \\
&\quad + \sup_{\hat{x}\in\sqrt{d}\,\mathbb{S}^{d-1}} \left|\tilde{\varphi}(x_m^\epsilon)^\top \varphi(\hat{x}) - \tilde{\varphi}(\hat{x})^\top \varphi(\hat{x})\right| \\
&\leq \max_{m\in[M]} \|\tilde{\varphi}(x_m^\epsilon)\|_2\, L\,\|V\|_{\mathsf{op}} \sup_{\hat{x}\in\sqrt{d}\,\mathbb{S}^{d-1}} \|x_m^\epsilon - \hat{x}\|_2 \\
&\quad + \sup_{\hat{x}\in\sqrt{d}\,\mathbb{S}^{d-1}} \|\varphi(\hat{x})\|_2\, \tilde{L}\,\|V\|_{\mathsf{op}} \sup_{\hat{x}\in\sqrt{d}\,\mathbb{S}^{d-1}} \|x_m^\epsilon - \hat{x}\|_2 \\
&= O\left(\sqrt{p}\,\frac{\sqrt{p}}{\sqrt{d}}\frac{1}{d^{3/2}}\right) = O\left(\frac{p}{d^2}\right).
\end{aligned} \tag{E.19}$$

Thus, merging (E.15), (E.16) and (E.19) yields

$$\sup_{\hat{x}\in\sqrt{d}\,\mathbb{S}^{d-1}} \left|\tilde{\varphi}(\hat{x})^\top \varphi(\hat{x}) - \tilde{\mu}^2 p\right| = O\left(\sqrt{pd}\log d + \frac{p}{d^2}\right), \tag{E.20}$$

with probability at least $1 - 2\exp(-c_8 d\log^2 d)$ over V, which proves (E.7).

Let us now consider a fixed $i \in [n]$ and $m \in [M]$. We have

$$\begin{aligned}\tilde{\varphi}(x_i)^\top \varphi(x_m^\epsilon) - \tilde{\varphi}(x_i)^\top \tilde{\varphi}(x_m^\epsilon) = &\sum_{k=1}^{p} \left(\tilde{\phi}(v_k^\top x_i)\phi(v_k^\top x_m^\epsilon) - \mathbb{E}_{v_k}\left[\tilde{\phi}(v_k^\top x_i)\phi(v_k^\top x_m^\epsilon)\right]\right) \\ &- \sum_{k=1}^{p} \left(\tilde{\phi}(v_k^\top x_i)\tilde{\phi}(v_k^\top x_m^\epsilon) - \mathbb{E}_{v_k}\left[\tilde{\phi}(v_k^\top x_i)\tilde{\phi}(v_k^\top x_m^\epsilon)\right]\right) \\ &+ p\,\mathbb{E}_v\left[\tilde{\phi}(v^\top x_i)\phi(v^\top x_m^\epsilon)\right] - p\,\mathbb{E}_v\left[\tilde{\phi}(v^\top x_i)\tilde{\phi}(v^\top x_m^\epsilon)\right].\end{aligned} \tag{E.21}$$

The last line is equal to 0, due to the Hermite coefficients of ϕ and $\tilde{\phi}$. The first two lines of the RHS can be bounded separately via Bernstein inequality as in (E.11), giving

$$\left|\tilde{\varphi}(x_i)^\top \varphi(x_m^\epsilon) - \tilde{\varphi}(x_i)^\top \tilde{\varphi}(x_m^\epsilon)\right| = O(\sqrt{p}\,t), \tag{E.22}$$

with probability at least $1 - 2\exp(-c_9 t^2)$ over V, for any $0 < t < p$. As done in (E.16), considering again $\epsilon = 1/d^2$, we have that the previous bound holds uniformly for every $i \in [n]$ and for every $m \in [M]$, after setting $t = \sqrt{d}\log d$, with probability at least $1 - 2\exp(-c_{10} d\log^2 d)$. Then, a similar argument as the one used in (E.19) yields

$$\sup_{\hat{x} \in \sqrt{d}\,\mathbb{S}^{d-1}} \left|\tilde{\varphi}(x_i)^\top \varphi(\hat{x}) - \tilde{\varphi}(x_i)^\top \tilde{\varphi}(\hat{x})\right| = O\left(\sqrt{pd}\log d + \frac{p}{d^2}\right), \tag{E.23}$$

with probability at least $1 - 2\exp(-c_{11} d\log^2 d)$. This proves (E.8). The proof of (E.9) is analogous and the argument is complete. □

Lemma E.2. *Let $\hat{x}$ be a generic row of $\hat{X}$ such that $\mathcal{L}(\hat{X}) = 0$ and $a \in \mathbb{R}^n$ be defined according to (E.4). Then, we have that, for any $i \in [n]$,*

$$\left|\tilde{\varphi}(x_i)^\top \tilde{\varphi}(\hat{x}) - \tilde{\mu}^2 p a_i\right| = O\left((\|a\|_2 + 1)\left(\sqrt{pd}\log d + \sqrt{pn}\log d + p\frac{\sqrt{n}\log^3 d}{d^{3/2}}\right)\right), \tag{E.24}$$

with probability at least $1 - 2\exp(-c\log^2 d)$ over V and X.

Proof. For any $i \in [n]$, (E.4) implies

$$\begin{aligned}\tilde{\varphi}(x_i)^\top \varphi(\hat{x}) = \tilde{\varphi}(x_i)^\top \Phi^\top a &= \sum_{j=1}^{n} \tilde{\varphi}(x_i)^\top \varphi(x_j) a_j \\ &= \tilde{\mu}^2 p a_i + \left(\tilde{\varphi}(x_i)^\top \varphi(x_i) a_i - \tilde{\mu}^2 p a_i\right) + \sum_{j \neq i} \tilde{\varphi}(x_i)^\top \varphi(x_j) a_j.\end{aligned} \tag{E.25}$$

This implies

$$\begin{aligned}\left|\tilde{\varphi}(x_i)^\top \varphi(\hat{x}) - \tilde{\mu}^2 p a_i\right| &\le \left|\tilde{\varphi}(x_i)^\top \varphi(x_i) - \tilde{\mu}^2 p\right| |a_i| + \sqrt{\sum_{j \neq i} \left(\tilde{\varphi}(x_i)^\top \varphi(x_j)\right)^2}\,\|a\|_2 \\ &= O\left(\|a\|_2 \left(\sqrt{p}\log d + \sqrt{n\left(p\log^2 d + p^2\frac{\log^6 d}{d^3}\right)}\right)\right) \\ &= O\left(\|a\|_2 \left(\sqrt{p}\log d + \sqrt{pn}\log d + p\frac{\sqrt{n}\log^3 d}{d^{3/2}}\right)\right),\end{aligned} \tag{E.26}$$

with probability at least $1 - 2\exp\left(-c_1 \log^2 d\right)$ over V and X, where the second step holds due to the first two equations in the statement of Lemma E.1. Then, due to the fourth equation in the statement of Lemma E.1, we have

$$\left|\tilde{\varphi}(x_i)^\top \varphi(\hat{x}) - \tilde{\varphi}(x_i)^\top \tilde{\varphi}(\hat{x})\right| = O\left(\sqrt{pd}\log d + \frac{p}{d^2}\right), \tag{E.27}$$

which, together with (E.26), gives the desired result. □

Lemma E.3. *We have that*

$$\left|\|a\|_2^2 - 1\right| = O\left(\sqrt{\frac{dn}{p}}\log d + \frac{\log^3 d}{\sqrt{d}}\right), \tag{E.28}$$

with probability at least $1 - 2\exp(-c\log^2 d)$.

Proof. (E.4) directly implies

$$\tilde{\varphi}(\hat{x})^\top \varphi(\hat{x}) = \tilde{\varphi}(\hat{x})^\top \Phi^\top a_i = \sum_{j=1}^{n} \tilde{\varphi}(\hat{x})^\top \varphi(x_j) a_j. \tag{E.29}$$

Due to the third and fifth equations in the statement of Lemma E.1, we respectively have that

$$\left|\tilde{\varphi}(\hat{x})^\top \varphi(\hat{x}) - \tilde{\mu}^2 p\right| = O\left(\sqrt{pd}\log d + \frac{p}{d^2}\right), \tag{E.30}$$

$$\left|\varphi(x_i)^\top \tilde{\varphi}(\hat{x}) - \tilde{\varphi}(x_i)^\top \tilde{\varphi}(\hat{x})\right| = O\left(\sqrt{pd}\log d + \frac{p}{d^2}\right), \tag{E.31}$$

with probability at least $1 - 2\exp(-c_1 d\log^2 d)$. Then, by Cauchy-Schwartz inequality, (E.31) yields

$$\left|\sum_{i=1}^{n} \tilde{\varphi}(\hat{x})^\top \varphi(x_i) a_i - \sum_{i=1}^{n} \tilde{\varphi}(\hat{x})^\top \tilde{\varphi}(x_i) a_i\right| = O\left(\|a\|_2 \left(\sqrt{pdn}\log d + \frac{p\sqrt{n}}{d^2}\right)\right). \tag{E.32}$$

Again by Cauchy-Schwartz inequality, we have

$$\begin{aligned}
&\left|\sum_{i=1}^{n} a_i \left(\tilde{\varphi}(\hat{x})^\top \tilde{\varphi}(x_i) - \tilde{\mu}^2 p a_i\right)\right| = \\
&= O\left(\|a\|_2 \left(\|a\|_2 + 1\right)\left(\sqrt{pnd}\log d + \sqrt{p}n\log d + pn\frac{\log^3 d}{d^{3/2}}\right)\right) \\
&= O\left(\|a\|_2 \left(\|a\|_2 + 1\right)\left(\sqrt{pnd}\log d + p\frac{\log^3 d}{\sqrt{d}}\right)\right),
\end{aligned} \tag{E.33}$$

with probability at least $1 - 2\exp(-c_2\log^2 d)$, due to Lemma E.2, where the last step is a consequence of Assumption 4.3. Then, merging (E.29), (E.30), (E.32) and (E.33), an application of the triangle inequality yields

$$\tilde{\mu}^2 p\left|\|a\|_2^2 - 1\right| = O\left(\left(\|a\|_2^2 + 1\right)\left(\sqrt{pdn}\log d + p\frac{\log^3 d}{\sqrt{d}}\right)\right). \tag{E.34}$$

Notice that, if $\|a\|_2 = \omega(1)$, the LHS of the previous equation would be $\Omega(p\|a\|_2^2)$, while the RHS would be $o(p\|a\|_2^2)$ due to Assumption 4.3. Then, we necessarily have that $\|a\|_2 = O(1)$, which yields

$$\left|\|a\|_2^2 - 1\right| = O\left(\sqrt{\frac{dn}{p}}\log d + \frac{\log^3 d}{\sqrt{d}}\right), \tag{E.35}$$

with probability at least $1 - 2\exp(-c_3\log^2 d)$, which gives the desired result. □

Lemma E.4. *Let $\hat{x} \in \sqrt{d}\,\mathbb{S}^{d-1}$ a generic vector, and let $0 \leq C \leq 1$ be defined as*

$$C = \max_i \left| \frac{x_i^\top \hat{x}}{d} \right|. \tag{E.36}$$

Then, with probability at least $1 - 2\exp\left(-c\sqrt{d}\right)$ over X, we have that

$$\left\| \frac{(X\hat{x})^{\circ l}}{d^l} \right\|_2 \leq C^{l-1} + O\left(d^{-0.1}\right), \tag{E.37}$$

uniformly for every $l \geq 2$.

Proof. Fix $0 \leq \delta \leq 1$, and consider the values of $i \in [n]$ such that

$$\left| \frac{x_i^\top \hat{x}}{d} \right| > \delta. \tag{E.38}$$

Define $M \subseteq [n]$ as the set containing all the values of i that satisfy the inequality above, $m = |M|$ as the cardinality of M, and $X_\delta \in \mathbb{R}^{m \times d}$ as the matrix that contains in its rows all and only the x_i-s such that $i \in M$. Note that $X^\top$ is a matrix with independent sub-Gaussian columns, with fixed ℓ_2 norm equal to $\sqrt{d}$. Then, Theorem 1.3 in [Plan and Vershynin, 2025] yields

$$\left| \|X\|_{\mathsf{op}} - \sqrt{d} \right| = O(\sqrt{n}), \tag{E.39}$$

with probability at least $1 - 2\exp(-c_1 n)$. This, due to Assumption 4.3, guarantees that $\left\|X/\sqrt{d}\right\|_{\mathsf{op}} = O(1)$, and therefore

$$\delta^2 m \leq \left\| \frac{X\hat{x}}{d} \right\|_2^2 \leq \left\| \frac{X}{\sqrt{d}} \right\|_{\mathsf{op}}^2 \left\| \frac{\hat{x}}{\sqrt{d}} \right\|_2^2 = O(1), \tag{E.40}$$

which implies

$$m = O\left(\frac{1}{\delta^2}\right). \tag{E.41}$$

Consider all the possible subsets of size m of $[n]$. There are in total

$$\binom{n}{m} \leq n^m \leq \exp\left(m \log n\right) \leq \exp\left(C_1 \frac{\log d}{\delta^2}\right) \tag{E.42}$$

such subsets, where C is an absolute constant, and where we used Assumption 4.3. For each of these subsets, the operator norm of the matrix $X_s \in \mathbb{R}^{m \times d}$ with rows indexed by M is

$$\left| \|X_s\|_{\mathsf{op}} - \sqrt{d} \right| \leq C_2 \left(\sqrt{m} + t\right), \tag{E.43}$$

with probability $1 - 2\exp(-c_2 t^2)$, again due to Theorem 1.3 in [Plan and Vershynin, 2025]. Then, performing a union bound over all subsets and setting $t = \sqrt{2C_1/c_2}\frac{\sqrt{\log d}}{\delta}$, we have that all matrices with rows belonging to a generic subset of size m of the rows of X respect (E.43) with probability at least $1 - 2\exp\left(-C_1 \frac{\log d}{\delta^2}\right)$. In particular, with this probability, we also have

$$\left| \|X_\delta\|_{\mathsf{op}} - \sqrt{d} \right| \leq C_2 \left(\sqrt{m} + t\right) = O\left(\frac{\sqrt{\log d}}{\delta}\right), \tag{E.44}$$

where in the second step we used (E.41). Then, we have that

$$\left\|\frac{(X\hat{x})^{\circ l}}{d^l}\right\|_2^2 \leq \left\|\frac{(X_\delta\hat{x})^{\circ l}}{d^l}\right\|_2^2 + (n-m)\delta^{2l} \leq C^{2l-2}\left\|\frac{X_\delta\hat{x}}{d}\right\|_2^2 + n\delta^4, \tag{E.45}$$

where in the last step we used that $l \geq 2$, and the definition of C in (E.36). Setting $\delta = d^{-0.3}$, Assumption 4.3 gives $n\delta^4 = O(d^{-0.2})$, which yields

$$\left\|\frac{X_\delta\hat{x}}{d}\right\|_2 \leq \left\|X_\delta/\sqrt{d}\right\|_{\mathsf{op}} \leq 1 + \left|\left\|X_\delta/\sqrt{d}\right\|_{\mathsf{op}} - 1\right| \leq 1 + C_3\frac{\sqrt{\log d}+1}{\sqrt{d}\,d^{-0.3}}, \tag{E.46}$$

where the last step holds due to (E.44) with probability at least $1 - 2\exp\left(-C_1\frac{\log d}{d^{-0.6}}\right) \geq 1 - 2\exp\left(-c_3\sqrt{d}\right)$. Thus, plugging in (E.45), the thesis readily follows. □

Proof of Theorem 4.4. As done in Lemma E.1, consider the $\sqrt{d}\,\epsilon$-net of the sphere $\sqrt{d}\,\mathbb{S}^{d-1}$, with $\epsilon = 1/d$. For any element x_m^ϵ of this set (with a fixed index $m \in [M]$, where M denotes the cardinality of the net), (E.4) yields

$$\tilde{\varphi}(x_m^\epsilon)^\top\varphi(\hat{x}) = \tilde{\varphi}(x_m^\epsilon)^\top\Phi^\top a = \sum_{i=1}^n \tilde{\varphi}(x_m^\epsilon)^\top\varphi(x_i)a_i, \tag{E.47}$$

where each term of the sum above reads

$$\begin{aligned}\tilde{\varphi}(x_m^\epsilon)^\top\varphi(x_i) = \sum_{k=1}^p &\left(\tilde{\phi}(v_k^\top x_m^\epsilon)\phi(v_k^\top x_i) - \mathbb{E}_{v_k}\left[\tilde{\phi}(v_k^\top x_m^\epsilon)\phi(v_k^\top x_i)\right]\right)\\ &+ p\,\mathbb{E}_v\left[\tilde{\phi}(v^\top x_m^\epsilon)^\top\phi(v^\top x_i)\right].\end{aligned} \tag{E.48}$$

By Bernstein inequality (see the same argument as in (E.11)), we have that

$$\sum_{k=1}^p\left(\tilde{\phi}(v_k^\top x_m^\epsilon)\phi(v_k^\top x_i) - \mathbb{E}_{v_k}\left[\tilde{\phi}(v_k^\top x_m^\epsilon)\phi(v_k^\top x_i)\right]\right) = O\left(\sqrt{pd}\log d\right), \tag{E.49}$$

with probability at least $1 - 2\exp\left(-c_1 d\log^2 d\right)$. Performing a union bound over the elements of the net, we have that (E.49) holds uniformly for all $i \in [n]$ and all $m \in [M]$ with probability at least $1 - 2\exp\left(-c_2 d\log^2 d\right)$ (see the argument prior to (E.16)). Furthermore, the Hermite decomposition of ϕ and $\tilde{\phi}$ gives

$$\mathbb{E}_v\left[\tilde{\phi}(v^\top x_m^\epsilon)\phi(v^\top x_i)\right] = \sum_{l=3}^{+\infty}\mu_l^2\frac{\left(x_i^\top x_m^\epsilon\right)^l}{d^l}, \tag{E.50}$$

which thus allows us to write

$$\begin{aligned}\left|\sum_{i=1}^n\tilde{\varphi}(x_m^\epsilon)^\top\varphi(x_i)a_i - p\sum_{i=1}^n a_i\sum_{l=3}^{+\infty}\mu_l^2\frac{\left(x_i^\top x_m^\epsilon\right)^l}{d^l}\right| &= O\left(\|a\|_2\sqrt{pdn}\log d\right)\\ &= O\left(\sqrt{pdn}\log d\right),\end{aligned} \tag{E.51}$$

where the first step follows from (E.48), (E.49) and (E.50), and an application of Cauchy Schwartz inequality; the second step holds due to Lemma E.3 with probability at least $1 - 2\exp\left(-c_3\log^2 d\right)$.

By the definition of the net, there exists $m \in [M]$ such that $\|x_m^\epsilon - \hat{x}\|_2 \le 1/\sqrt{d}$. Fixing such m, we have

$$\begin{aligned}\left|\tilde{\mu}^2 p - p\sum_{i=1}^n a_i \sum_{l=3}^{+\infty} \mu_l^2 \frac{\left(x_i^\top x_m^\epsilon\right)^l}{d^l}\right| &\le \left|\tilde{\mu}^2 p - \tilde{\varphi}(\hat{x})^\top \varphi(\hat{x})\right| + \left|\tilde{\varphi}(\hat{x})^\top \varphi(\hat{x}) - \tilde{\varphi}(x_m^\epsilon)^\top \varphi(\hat{x})\right| \\ &\quad + \left|\sum_{i=1}^n \tilde{\varphi}(x_m^\epsilon)^\top \varphi(x_i) a_i - p\sum_{i=1}^n a_i \sum_{l=3}^{+\infty} \mu_l^2 \frac{\left(x_i^\top x_m^\epsilon\right)^l}{d^l}\right| \\ &= O\left(\sqrt{pd}\log d + \frac{p}{d} + \frac{p}{d} + \sqrt{pdn}\log d\right) \\ &= O\left(\sqrt{pdn}\log d + \frac{p}{d}\right),\end{aligned} \tag{E.52}$$

due to the third equation in the statement of Lemma E.1, an argument equivalent to the one in (E.19), and (E.51). Considering the union bound on these high probability events, we have that (E.52) holds with probability at least $1 - 2\exp\left(-c_4 \log^2 d\right)$.

Let's suppose we have

$$\max_i \left|\frac{\hat{x}^\top x_i}{d}\right| \le C < 1, \tag{E.53}$$

where C is an absolute constant (independent of d, n, p). Thus,

$$\max_i \left|\frac{{x_m^\epsilon}^\top x_i}{d}\right| \le \max_i \left|\frac{\hat{x}^\top x_i}{d}\right| + \max_i \left|\frac{\|x_m^\epsilon - \hat{x}\|_2 \|x_i\|_2}{d}\right| \le C + \frac{1}{d} \le C_\epsilon < 1, \tag{E.54}$$

where the last inequality holds for sufficiently large d. Then, we have

$$\begin{aligned}\left|\sum_{i=1}^n a_i \sum_{l=3}^{+\infty} \mu_l^2 \frac{\left(x_i^\top x_m^\epsilon\right)^l}{d^l}\right| &\le \|a\|_2 \sum_{l=3}^{+\infty} \mu_l^2 \left\|\frac{(X x_m^\epsilon)^{\circ l}}{d^l}\right\|_2 \\ &\le \left(1 + C_1\left(\sqrt{\frac{dn}{p}}\log d + \frac{\log^3 d}{\sqrt{d}}\right)\right) \sum_{l=3}^{+\infty} \mu_l^2 \left(C_\epsilon^{l-1} + C_2 d^{-0.1}\right) \\ &\le \tilde{\mu}^2 C_\epsilon^2 + C_3\left(\sqrt{\frac{dn}{p}}\log d + d^{-0.1}\right),\end{aligned} \tag{E.55}$$

where we applied Cauchy Schwartz inequality separately for every l in the first step and used Lemmas E.3 and E.4 in the second step. Here, C_1 and C_2 denote two positive absolute constants, and the inequality holds with probability at least $1 - 2\exp\left(-c_5 \log^2 d\right)$. Plugging (E.55) in (E.52) yields

$$\tilde{\mu}^2\left(1 - C_\epsilon^2\right) = O\left(\sqrt{\frac{dn}{p}}\log d + d^{-0.1}\right) = o(1), \tag{E.56}$$

where the last step is a consequence of Assumption 4.3, which provides a contradiction with the hypothesis in (E.53), implying that

$$\left|1 - \max_i \left|\frac{\hat{x}^\top x_i}{d}\right|\right| = o(1), \tag{E.57}$$

with probability at least $1 - 2\exp\left(-c_5 \log^2 d\right)$.

Let $j = \arg\max_i \left|\hat{x}^\top x_i\right|$ (taking the smallest index if there are multiple indices that maximize this value), and suppose $\hat{x}^\top x_i \geq 0$. Then, the law of cosines yields

$$\|\hat{x} - x_j\|_2 = \sqrt{2d\left(1 - \frac{\hat{x}^\top x_j}{d}\right)} = o\left(\sqrt{d}\right), \tag{E.58}$$

where the last step holds due to (E.57). Thus, for $k \neq j$ we have

$$\left|\frac{x_k^\top \hat{x}}{d}\right| = \left|\frac{x_k^\top x_j}{d} + \frac{x_k^\top (\hat{x} - x_j)}{d}\right| \leq \frac{\left|x_k^\top x_j\right|}{d} + \frac{\|x_k\|_2 \|\hat{x} - x_j\|_2}{d} = o(1), \tag{E.59}$$

where the last step holds with probability at least $1 - 2\exp\left(-c_6 \log^2 d\right)$ due to (E.58) and (E.13). In the case $\hat{x}^\top x_i < 0$ the same argument with $-x_j$ instead of x_j yields the same thesis. Thus, comparing with (E.57), we have that $\arg\max_i \left|\hat{x}^\top x_i\right|$ has a unique solution j, and all other indices are such that $\left|\hat{x}^\top x_i\right| / d = o(1)$.

This proves that $\hat{x}$ is aligned with x_i. It remains to prove that $\hat{x}$ also has the correct sign (i.e., it is close to x_i and not to $-x_i$). To do so, following the same approach that led to (E.55), we have that

$$\left|\sum_{i \neq j} a_i \sum_{l=3}^{+\infty} \mu_l^2 \frac{\left(x_i^\top x_m^\epsilon\right)^l}{d^l}\right| = o(1), \tag{E.60}$$

with probability at least $1 - 2\exp\left(-c_7 \log^2 d\right)$. Thus, comparing with (E.52), we get

$$\left|\tilde{\mu}^2 - a_j \sum_{l=3}^{+\infty} \mu_l^2 \frac{\left(x_j^\top x_m^\epsilon\right)^l}{d^l}\right| = o(1). \tag{E.61}$$

Since $|a_j| \leq 1 + o(1)$ by Lemma E.3, (E.61) can hold only if $\left(x_j^\top x_m^\epsilon\right)^l$ have all the same sign for all $l \geq 3$ such that $\mu_l \neq 0$. By Assumption 4.2, this is possible only if $x_j^\top x_m^\epsilon > 0$, which concludes the argument. □

E.2 Proof of Theorem 4.5

Due to Theorem 4.4, we have that, with overwhelming probability, every $\hat{x}_{\hat{\imath}}$ has a unique "closest" row vector in X, such that

$$\left|1 - \frac{\hat{x}_{\hat{\imath}}^\top x_i}{d}\right| = o(1). \tag{E.62}$$

Then, if we consider the case $n = 2$, either both samples are reconstructed, or the same training sample is reconstructed twice. By contradiction, let us suppose the latter hypothesis, which without loss of generality can be framed as the first sample x_1 being reconstructed twice.

Since we have that $\varphi(x_2) \in \mathrm{Span}\{\mathrm{rows}(\hat{\Phi})\}$, which means that there exist two real numbers a_1 and a_2 such that

$$\varphi(x_2) = a_1\varphi(x_1+\varepsilon_1) + a_2\varphi(x_1+\varepsilon_2), \tag{E.63}$$

where

$$\|x_1+\varepsilon_1\|_2 = \|x_1+\varepsilon_2\|_2 = \sqrt{d}, \tag{E.64}$$

as we are considering data reconstruction on the sphere. From now on, we will always assume that (E.63) and (E.64) hold. We will often consider the following expansion:

$$\begin{aligned}\varphi(x_2) &= a_1\varphi(x_1+\varepsilon_1) + a_2\varphi(x_1+\varepsilon_2)\\ &= a_1\varphi(x_1+\varepsilon_1) + a_2\varphi(x_1+\varepsilon_1+(\varepsilon_2-\varepsilon_1))\\ &= (a_1+a_2)\varphi(x_1+\varepsilon_1)\\ &\quad + a_2\left((\varphi(x_1+\varepsilon_2)-\varphi(x_1+\varepsilon_1)) - \phi'(V(x_1+\varepsilon_1))\circ(V(\varepsilon_2-\varepsilon_1))\right)\\ &\quad + a_2\,\phi'(V(x_1+\varepsilon_1))\circ(V(\varepsilon_2-\varepsilon_1))\,,\end{aligned} \tag{E.65}$$

which can be used since ϕ admits first derivative according to Assumption 4.2. Note that

$$0 = \|x_1+\varepsilon_1\|_2^2 - \|x_1\|_2^2 = 2x_1^\top\varepsilon_1 + \|\varepsilon_1\|_2^2\,, \tag{E.66}$$

which yields

$$2\left|x_1^\top\varepsilon_1\right| = \|\varepsilon_1\|_2^2\,, \tag{E.67}$$

with the same relation holding for ε_2. Similarly, we have

$$0 = \|x_1+\varepsilon_1\|_2^2 - \|x_1+\varepsilon_2\|_2^2 = 2x_1^\top(\varepsilon_1-\varepsilon_2) + (\|\varepsilon_1\|_2+\|\varepsilon_2\|_2)(\|\varepsilon_1\|_2-\|\varepsilon_2\|_2)\,, \tag{E.68}$$

which yields

$$2\left|x_1^\top(\varepsilon_1-\varepsilon_2)\right| \le (\|\varepsilon_1\|_2+\|\varepsilon_2\|_2)\,\|\varepsilon_1-\varepsilon_2\|_2\,. \tag{E.69}$$

We will use the following notation

$$\varepsilon = \frac{\max\{\|\varepsilon_1\|_2, \|\varepsilon_2\|_2\}}{\sqrt{d}} = O(1), \tag{E.70}$$

and

$$\delta = \frac{\|\varepsilon_2-\varepsilon_1\|_2}{\sqrt{d}} = O(1). \tag{E.71}$$

Notice that, by definition, we have $\delta = O(\epsilon)$. The idea is to prove that, with overwhelming probability, there exists no solution to (E.63) such that $\varepsilon = o(1)$. This then readily implies the claim of Theorem 4.5. To do so, we state and prove a number of preliminary results.

Lemma E.5. *We jointly have*

$$\|V\|_{\mathsf{op}} = O(\sqrt{p/d}), \qquad \|V\|_F^2 = O(p)\,, \qquad \sum_{k=1}^{p}\|v_k\|_2^3 = O(p)\,, \tag{E.72}$$

with probability at least $1-2\exp\left(-c\log^2 d\right)$ *over* V*. Furthermore, we have that*

$$\sup_{x\in\sqrt{d}\,\mathbb{S}^{d-1}}\|\varphi(x)\|_2 = \Theta(\sqrt{p}), \qquad \sup_{x\in\sqrt{d}\,\mathbb{S}^{d-1}}\|\tilde{\varphi}(x)\|_2 = \Theta(\sqrt{p}). \tag{E.73}$$

with probability at least $1-2\exp\left(-c\log^2 d\right)$ *over* V*.*

Proof. The first equation holds with probability at least $1 - 2\exp(-c_1 p)$ due to Theorem 4.4.5 in [Vershynin, 2018]. The second equation is a direct consequence of Theorem 3.1.1 in [Vershynin, 2018]. For the third equation, due to Theorem 3.1.1 in [Vershynin, 2018], we have that

$$\left\| \|v_k\|_2 - 1 \right\|_{\psi_2} = O\left(\frac{1}{\sqrt{d}}\right), \tag{E.74}$$

which implies

$$\left\| \|v_k\|_2 \right\|_{\psi_2} = O\left(1\right). \tag{E.75}$$

Then, we have that the Orlicz norm $\left\| \|v_k\|_2^3 \right\|_{\psi_{2/3}} = O(1)$, which also implies $\mathbb{E}\left[\|v_k\|_2^3\right] = O(1)$. Then, due to Lemma B.6 in [Bombari et al., 2023], we have that

$$\sum_{k=1}^{p} \|v_k\|_2^3 = O(p), \tag{E.76}$$

with probability at least $1 - 2\exp\left(-c_1 \log^2 d\right)$, where we also used Assumption 4.3. The last statement statement can be obtained via the same argument in (E.18). □

Lemma E.6. *Let $\rho_1, \rho_2, \rho_3 \in \mathbb{R}$ be sub-Gaussian random variables, not necessarily independent. Consider the random variable*

$$Z = \min\left(M, |\rho_1|\right) |\rho_2 \rho_3|, \tag{E.77}$$

where $M = \omega(1)$. Then, $Z - \mathbb{E}[Z]$ is sub-exponential with parameters (ν, α) (see Definition 2.7 in [Wainwright, 2019]) such that

$$\nu = O(1), \qquad \alpha = O(M). \tag{E.78}$$

Proof. Since the ρ-s are sub-Gaussian, we have that both the absolute value of the mean and the second moment of Z (and therefore its variance) are upper bounded by positive absolute constants. Denote with C_1 a constant order upper bound on the absolute value of $\mathbb{E}[Z]$. Furthermore, without loss of generality, we will consider the sub-Gaussian norms of ρ_1, ρ_2, ρ_3 to be equal to 1.

Consider a real number λ such that $4M|\lambda| \leq 1$. Defining $\bar{Z} = Z - \mathbb{E}[Z]$, and noting that $e^z \leq 1 + z + z^2 e^{|z|}/2$ for every $z \in \mathbb{R}$ (due to the mean-value theorem), we have that

$$\begin{aligned}
\mathbb{E}\left[e^{\lambda \bar{Z}}\right] &\leq \mathbb{E}\left[1 + \lambda \bar{Z} + \frac{\lambda^2}{2} \bar{Z}^2 e^{|\lambda||\bar{Z}|}\right] \\
&\leq 1 + \mathbb{E}\left[\frac{\lambda^2}{2} \bar{Z}^2 e^{|\lambda||\bar{Z}|}\right] \\
&\leq 1 + \frac{\lambda^2}{2} \mathbb{E}\left[\bar{Z}^2 e^{\frac{M|\rho_2\rho_3| + C_1}{4M}}\right] \\
&\leq 1 + \frac{\lambda^2}{2} \mathbb{E}\left[\left(|\rho_1\rho_2\rho_3| + C_1\right)^2 e^{\frac{|\rho_2\rho_3|+1}{4}}\right] \\
&\leq 1 + \frac{\lambda^2}{2} \mathbb{E}\left[\left(|\rho_1\rho_2\rho_3| + C_1\right)^4\right]^{1/2} \mathbb{E}\left[e^{\frac{|\rho_2\rho_3|+1}{2}}\right]^{1/2} \\
&\leq 1 + C_2 \lambda^2 \\
&\leq e^{C_2 \lambda^2}.
\end{aligned} \tag{E.79}$$

Here, in third line we used $Z \leq M|\rho_2\rho_3|$; in the fourth line we used that $M \geq C_1$ and $Z \leq |\rho_1\rho_2\rho_3|$; in the sixth line we upper bounded the expectations via an absolute constant, as $|\rho_2\rho_3|$ is sub-exponential with norm 1. Thus, we can set $\alpha = 4M$, and for all $|\lambda| \leq 1/\alpha$, we have

$$\mathbb{E}\left[e^{\lambda \bar{Z}}\right] \leq e^{C_2\lambda^2}, \tag{E.80}$$

which gives the desired result according to Definition 2.7 in [Wainwright, 2019]. □

Lemma E.7. *We have that*

$$\begin{aligned}&\left|(\varepsilon_2-\varepsilon_1)^\top V^\top\left((\varphi(x_1+\varepsilon_2)-\varphi(x_1+\varepsilon_1))-\phi'(V(x_1+\varepsilon_1))\circ(V(\varepsilon_2-\varepsilon_1))\right)\right| \\ &\quad = O\left(p\delta^3 + d\log^2 d\delta^2\right),\end{aligned} \tag{E.81}$$

with probability at least $1-2\exp\left(-c_6\log^2 d\right)$ *over* V.

Proof. Since ϕ is differentiable due to Assumption 4.2, by the mean-value theorem, we have that there exists $\zeta \in \mathbb{R}^p$ such that

$$\varphi(x_1+\varepsilon_2)-\varphi(x_1+\varepsilon_1) = \phi'(V(x_1+\varepsilon_1)+\zeta)\circ(V(\varepsilon_2-\varepsilon_1)), \tag{E.82}$$

where each entry of ζ is such that $\zeta_k \in [0, [V(\varepsilon_2-\varepsilon_1)]_k]$ (or $\zeta_k \in [[V(\varepsilon_2-\varepsilon_1)]_k, 0]$, if $[V(\varepsilon_2-\varepsilon_1)]_k$ is negative). Then, we have that the thesis becomes

$$\begin{aligned}&\left|(\varepsilon_2-\varepsilon_1)^\top V^\top\left((\phi'(V(x_1+\varepsilon_1)+\zeta)-\phi'(V(x_1+\varepsilon_1)))\circ(V(\varepsilon_2-\varepsilon_1))\right)\right| \\ &\quad = O\left(p\delta^3 + d\log^2 d\delta^2\right).\end{aligned} \tag{E.83}$$

First, let $\xi \in \mathbb{R}^d$ be defined as

$$\xi = \frac{\varepsilon_2-\varepsilon_1}{\delta}, \tag{E.84}$$

i.e. the vector lying on the sphere $\sqrt{d}\,\mathbb{S}^{d-1}$ with the same direction as $\varepsilon_2-\varepsilon_1$. Then, dividing both sides of (E.83) by δ^3 and expanding the sum, the desired result can be reformulated as

$$\sum_{k=1}^p \left(v_k^\top\xi\right)\frac{\phi'(v_k^\top(x_1+\varepsilon_1)+\zeta_k)-\phi'(v_k^\top(x_1+\varepsilon_1))}{\delta}\left(v_k^\top\xi\right) = O\left(p+\frac{d\log^2 d}{\delta}\right). \tag{E.85}$$

Due to Assumption 4.2, we have both $\phi'(z) \leq L$ and $|\phi'(z_1)-\phi'(z_2)| \leq L'|z_1-z_2|$. Thus, the previous equation is implied by

$$\sum_{k=1}^p \min\left(\frac{2L}{\delta}, L'\left|v_k^\top\xi\right|\right)\left(v_k^\top\xi\right)^2 = O\left(p+\frac{d\log^2 d}{\delta}\right), \tag{E.86}$$

as we have $|\zeta_k| \leq \delta|v_k^\top\xi|$ from its definition in (E.82).

Let us now consider an $\epsilon\sqrt{d}$-net of $\sqrt{d}\,\mathbb{S}^{d-1}$, namely $\{x_m^\epsilon\}_{m=1}^M$, such that for any $x \in \sqrt{d}\,\mathbb{S}^{d-1}$ there exists $m \in [M]$ such that $\|x-x_m^\epsilon\|_2 \leq \epsilon\sqrt{d}$. Due to Corollary 4.2.13 in [Vershynin, 2018], for $\epsilon < 1$ we have that the net can be chosen such that $M \leq (3/\epsilon)^d$. Setting $\epsilon = d^{-3/2}$, we have that there exists $m^* \in [M]$ such that $\|\xi - x_{m^*}^\epsilon\|_2 \leq 1/d$, and $M \leq e^{c_1 d\log d}$, where c_1 is an absolute positive constant. Consider a generic element x_m^ϵ and define

$$T^{(m)} = \sum_{k=1}^p \min\left(\frac{2L}{\delta}, L'\left|v_k^\top x_m^\epsilon\right|\right)\left(v_k^\top x_m^\epsilon\right)^2. \tag{E.87}$$

Each term $T_k^{(m)}$ in the sum above is such that its expectation is

$$\mathbb{E}_{v_k}\left[T_k^{(m)}\right] = \mathbb{E}_{v_k}\left[\min\left(\frac{2L}{\delta}, L'\left|v_k^\top x_m^\epsilon\right|\right)\left(v_k^\top x_m^\epsilon\right)^2\right] \leq L'\mathbb{E}_{v_k}\left[\left|v_k^\top x_m^\epsilon\right|^3\right] = O(1). \quad \text{(E.88)}$$

Furthermore, $T_k^{(m)}$ is sub-exponential (in the probability space of v_k) with parameters (ν, α) (see Definition 2.7 in [Wainwright, 2019]) such that

$$\nu = O(1), \qquad \alpha = O\left(\delta^{-1}\right), \quad \text{(E.89)}$$

due to Lemma E.6. Thus, Equation (2.18) in [Wainwright, 2019] guarantees that

$$\begin{aligned}\mathbb{P}_V\left(\left|\sum_{k=1}^p\left(T_k^{(m)} - \mathbb{E}_{v_k}\left[T_k^{(m)}\right]\right)\right| \geq p + \frac{d\log^2 d}{\delta}\right) &\leq \exp\left(-c_2\min\left(p, d\log^2 d\right)\right) \\ &\leq \exp\left(-c_3 d\log^2 d\right),\end{aligned} \quad \text{(E.90)}$$

where the last step used Assumption 4.3. Next, we perform a union bound on the elements of the net, obtaining that

$$\sup_{m\in[M]}\left|T^{(m)}\right| = O\left(p + \frac{d\log^2 d}{\delta}\right), \quad \text{(E.91)}$$

with probability at least $1 - \exp\left(-c_3 d\log^2 d + c_1 d\log d\right) \geq 1 - 2\exp\left(-c_4 d\log^2 d\right)$. Then, with this same probability, due to (E.88), we also have that

$$\sum_{k=1}^p \min\left(\frac{2L}{\delta}, L'\left|v_k^\top x_{m^*}^\epsilon\right|\right)\left(v_k^\top x_{m^*}^\epsilon\right)^2 = O\left(p + \frac{d\log^2 d}{\delta}\right). \quad \text{(E.92)}$$

Since the $\min$ function is 1 Lipschitz in any of its arguments, for every $k \in [p]$, we have

$$\min\left(\frac{2L}{\delta}, L'\left|v_k^\top x_{m^*}^\epsilon\right|\right) - \min\left(\frac{2L}{\delta}, L'\left|v_k^\top \xi\right|\right) \leq L'\left\|v_k\right\|_2\left\|\xi - x_{m^*}^\epsilon\right\|_2 = O\left(\frac{\left\|v_k\right\|_2}{d}\right). \quad \text{(E.93)}$$

Thus,

$$\begin{aligned}&\sum_{k=1}^p\left|\min\left(\frac{2L}{\delta}, L'\left|v_k^\top x_{m^*}^\epsilon\right|\right) - \min\left(\frac{2L}{\delta}, L'\left|v_k^\top \xi\right|\right)\right|\left(v_k^\top x_{m^*}^\epsilon\right)^2 \\ &\qquad \leq \sum_{k=1}^p \frac{C_2}{d}\left\|v_k\right\|_2^3\left\|x_{m^*}^\epsilon\right\|_2^2 = O(p),\end{aligned} \quad \text{(E.94)}$$

where the last step used the first statement of Lemma E.5, and holds with probability at least $1 - 2\exp\left(-c_5\log^2 d\right)$. Furthermore, for every $k \in [p]$, we have

$$\begin{aligned}\left|\left(v_k^\top x_{m^*}^\epsilon\right)^2 - \left(v_k^\top \xi\right)^2\right| &= \left|v_k^\top\left(x_{m^*}^\epsilon - \xi\right)\right|\left|v_k^\top\left(x_{m^*}^\epsilon + \xi\right)\right| \\ &\leq 2\sqrt{d}\left\|v_k\right\|_2^2\left\|x_{m^*}^\epsilon - \xi\right\|_2 = O\left(\frac{\left\|v_k\right\|_2^2}{\sqrt{d}}\right),\end{aligned} \quad \text{(E.95)}$$

which yields

$$\sum_{k=1}^p \min\left(\frac{2L}{\delta}, L'\left|v_k^\top \xi\right|\right)\left|\left(v_k^\top x_{m^*}^\epsilon\right)^2 - \left(v_k^\top \xi\right)^2\right| \leq \sum_{k=1}^p \frac{C_3}{\sqrt{d}}\left\|v_k\right\|_2^3\left\|\xi\right\|_2 = O(p), \quad \text{(E.96)}$$

again due to Lemma E.5. Finally, applying the triangle inequality to (E.92), (E.94), and (E.96), gives

$$\sum_{k=1}^{p} \min\left(\frac{2L}{\delta}, L'\left|v_k^\top \xi\right|\right)\left(v_k^\top \xi\right)^2 = O\left(p + \frac{d\log^2 d}{\delta}\right), \tag{E.97}$$

with probability at least $1 - 2\exp\left(-c_6 \log^2 d\right)$, which concludes the proof. □

Lemma E.8. *Suppose $\varepsilon = o(1)$. Then, we have that*

$$\left|(\varepsilon_2 - \varepsilon_1)^\top V^\top \left(\phi'(V(x_1+\varepsilon_1)) \circ (V(\varepsilon_2 - \varepsilon_1))\right) - \frac{\|\varepsilon_2 - \varepsilon_1\|_2^2}{d} p\mu_1\right| = o(p\delta^2), \tag{E.98}$$

with probability at least $1 - 2\exp\left(-c\log^2 d\right)$ over V.

Proof. First, let $\xi \in \mathbb{R}^d$ be defined as in (E.84). Then, the desired result can be reformulated as

$$\left|\sum_{k=1}^{p} \phi'\left(v_k^\top (x_1+\varepsilon_1)\right)\left(v_k^\top \xi\right)^2 - p\mu_1\right| = o(p). \tag{E.99}$$

Let us now consider an $\epsilon\sqrt{d}$-net of $\sqrt{d}\,\mathbb{S}^{d-1}$ (as we did after (E.86)). Setting $\epsilon = d^{-2}$, we have that there exist $m^{(\xi)}$ and $m^{(x)}$ in $[M]$ such that $\left\|\xi - x^\epsilon_{m^{(\xi)}}\right\|_2 \le 1/d^{3/2}$, $\left\|(x_1+\varepsilon_1) - x^\epsilon_{m^{(x)}}\right\|_2 \le 1/d^{3/2}$, and the size of the net is $M \le e^{c_1 d\log d}$, where c_1 is an absolute positive constant. Consider two generic elements of this net: $x^\epsilon_{m^{(1)}}$ and $x^\epsilon_{m^{(2)}}$, and define

$$T^{(m^{(1)},m^{(2)})} = \left|\sum_{k=1}^{p}\left(\phi'\left(v_k^\top x^\epsilon_{m^{(1)}}\right)\left(v_k^\top x^\epsilon_{m^{(2)}}\right)^2 - \mathbb{E}_{v_k}\left[\phi'\left(v_k^\top x^\epsilon_{m^{(1)}}\right)\left(v_k^\top x^\epsilon_{m^{(2)}}\right)^2\right]\right)\right|. \tag{E.100}$$

Since $\left|\phi'\left(v_k^\top x^\epsilon_{m^{(1)}}\right)\right| \le L$, each element of the sum is sub-exponential (in the probability space of v_k). Then, due to Bernstein inequality (see Theorem 2.8.1 in [Vershynin, 2018]), we have that

$$\begin{aligned}\mathbb{P}_V\left(T^{(m^{(1)},m^{(2)})} > C_1\sqrt{dp\log d}\right) &\le 2\exp\left(-c_2\min\left(C_1^2 d\log d, C_1\sqrt{dp\log d}\right)\right)\\ &\le 2\exp\left(-c_3 C_1 d\log d\right),\end{aligned} \tag{E.101}$$

where the second step is a consequence of Assumption 4.3. Performing a union bound, we have that, for every pair of points on the net (there are less than $e^{2c_1 d\log d}$ such pairs), including $x^\epsilon_{m^{(\xi)}}$ and $x^\epsilon_{m^{(x)}}$, the equation above holds with probability at least $1 - 2\exp\left(-c_3 C_1 d\log d + 2c_1 d\log d\right) \ge 1 - 2\exp\left(-c_4 d\log d\right)$ if C_1 is chosen sufficiently large. Thus, with this probability, we have

$$\begin{aligned}T^{(m^{(x)},m^{(\xi)})} &= \left|\sum_{k=1}^{p}\left(\phi'\left(v_k^\top x^\epsilon_{m^{(x)}}\right)\left(v_k^\top x^\epsilon_{m^{(\xi)}}\right)^2 - \mathbb{E}_{v_k}\left[\phi'\left(v_k^\top x^\epsilon_{m^{(x)}}\right)\left(v_k^\top x^\epsilon_{m^{(\xi)}}\right)^2\right]\right)\right|\\ &= \left|\sum_{k=1}^{p}\phi'\left(v_k^\top x^\epsilon_{m^{(x)}}\right)\left(v_k^\top x^\epsilon_{m^{(\xi)}}\right)^2 - p\,\mathbb{E}\left[\phi'\left(\rho_x^\epsilon\right)\left(\rho_\xi^\epsilon\right)^2\right]\right|\\ &= O\left(\sqrt{dp\log d}\right) = o(p),\end{aligned} \tag{E.102}$$

where we used Assumption 4.3 and introduced ρ_x^ϵ and ρ_ξ^ϵ as two standard Gaussian variables with correlation $\left(x^\epsilon_{m^{(x)}}\right)^\top x^\epsilon_{m^{(\xi)}}/d$. Note that

$$\left(\rho_\xi^\epsilon\right)^2 = 1 + \sqrt{2}\frac{\left(\rho_\xi^\epsilon\right)^2 - 1}{\sqrt{2}} = h_0\left(\rho_\xi^\epsilon\right) + \sqrt{2}h_2\left(\rho_\xi^\epsilon\right), \tag{E.103}$$

where h_0 and h_2 denote the 0th and 2nd Hermite polynomials. Thus, we have that

$$\mathbb{E}\left[\phi'\left(\rho_x^\epsilon\right)\left(\rho_\xi^\epsilon\right)^2\right] = \mu_0^{\phi'} + \sqrt{2}\mu_2^{\phi'}\left(\frac{\left(x_{m^{(x)}}^\epsilon\right)^\top x_{m^{(\xi)}}^\epsilon}{d}\right)^2, \tag{E.104}$$

where $\mu_0^{\phi'}$ is the 0th Hermite coefficient of ϕ'. Note that $\mu_0^{\phi'}$ corresponds to the 1st Hermite coefficient of ϕ, which is $\mu_1 \neq 0$ due to Assumption 4.2. The second term on the RHS of (E.104) can be bounded via

$$\begin{aligned}\left|\left(x_{m^{(x)}}^\epsilon\right)^\top x_{m^{(\xi)}}^\epsilon\right| &\leq \left\|x_{m^{(x)}}^\epsilon\right\|_2 \left\|x_{m^{(\xi)}}^\epsilon - \xi\right\|_2 + \left\|x_{m^{(x)}}^\epsilon - x_1 - \varepsilon_1\right\|_2 \|\xi\|_2 + \left|\left(x_1+\varepsilon_1\right)^\top \xi\right| \\ &\leq \frac{2}{d} + \left|\frac{x_1^\top\left(\varepsilon_1-\varepsilon_2\right)}{\delta}\right| + \left|\frac{\varepsilon_1^\top\left(\varepsilon_1-\varepsilon_2\right)}{\delta}\right| = O\left(\frac{1}{d}+d\varepsilon\right),\end{aligned} \tag{E.105}$$

where we used (E.69) in the last step. Then, plugging in (E.104), we obtain

$$\left|\mathbb{E}\left[\phi'\left(\rho_x^\epsilon\right)\left(\rho_\xi^\epsilon\right)^2\right] - \mu_1\right| = O\left(\frac{1}{d^4}+\varepsilon^2\right) = o(1). \tag{E.106}$$

We also have

$$\begin{aligned}&\left|\phi'\left(v_k^\top x_{m^{(x)}}^\epsilon\right)\left(v_k^\top x_{m^{(\xi)}}^\epsilon\right)^2 - \phi'\left(v_k^\top\left(x_1+\varepsilon_1\right)\right)\left(v_k^\top\xi\right)^2\right| \\ &\leq \left|\left(\phi'\left(v_k^\top x_{m^{(x)}}^\epsilon\right) - \phi'\left(v_k^\top\left(x_1+\varepsilon_1\right)\right)\right)\left(v_k^\top x_{m^{(\xi)}}^\epsilon\right)^2\right| \\ &\quad + \left|\phi'\left(v_k^\top\left(x_1+\varepsilon_1\right)\right)\left(\left(v_k^\top x_{m^{(\xi)}}^\epsilon\right)^2 - \left(v_k^\top\xi\right)^2\right)\right| \\ &\leq L'\|v_k\|_2\left\|x_{m^{(x)}}^\epsilon - x_1 - \varepsilon_1\right\|_2\|v_k\|_2^2\, d + 2L\sqrt{d}\,\|v_k\|_2^2\left\|x_{m^{(\xi)}}^\epsilon - \xi\right\|_2 \\ &\leq C_2\frac{\|v_k\|_2^2 + \|v_k\|_2^3}{\sqrt{d}},\end{aligned} \tag{E.107}$$

where C_2 is some positive constant. This yields

$$\sum_{k=1}^p\left|\phi'\left(v_k^\top x_{m^{(x)}}^\epsilon\right)\left(v_k^\top x_{m^{(\xi)}}^\epsilon\right)^2 - \phi'\left(v_k^\top\left(x_1+\varepsilon_1\right)\right)\left(v_k^\top\xi\right)^2\right| = O\left(\frac{p}{\sqrt{d}}\right) = o(p), \tag{E.108}$$

with probability at least $1-2\exp\left(-c_5\log^2 d\right)$ due to Lemma E.5. Then, applying the triangle inequality to (E.102), (E.106), and (E.108), the proof is complete. □

Lemma E.9. *Suppose $\varepsilon = o(1)$. Then, we have that*

$$|a_1+a_2| = O\left(|a_2|\,\delta + \frac{\log d}{\sqrt{d}}\right), \tag{E.109}$$

with probability at least $1-2\exp\left(-c\log^2 d\right)$ over V and X.

Proof. Consider (E.65), written as

$$\varphi(x_2) = (a_1+a_2)\varphi(x_1+\varepsilon_1) + a_2\left(\varphi(x_1+\varepsilon_2) - \varphi(x_1+\varepsilon_1)\right). \tag{E.110}$$

Let us now take an inner product of both sides of (E.110) with Vx_1. Due to Lemma E.5 and since x_1 is a sub-Gaussian vector independent of V and x_2 due to Assumption 4.1, we have that

$$\left|x_1^\top V^\top \varphi(x_2)\right| = O\left(\frac{p}{\sqrt{d}}\log d\right), \tag{E.111}$$

with probability at least $1-2\exp\left(-c_1\log^2 d\right)$ over x_1 and V. Then, consider

$$\begin{aligned}\left|x_1^\top V^\top \varphi(x_1+\varepsilon_1)-\mu_1 p\right| \leq &\left|x_1^\top V^\top \phi\left(V(x_1+\varepsilon_1)\right)-x_1^\top V^\top \phi(Vx_1)\right| \\ &+\left|x_1^\top V^\top \phi(Vx_1)-\mu_1 p\right|,\end{aligned} \tag{E.112}$$

and let us bound the two terms on the RHS separately. For the first one, since ϕ is L-Lipschitz by Assumption 4.2, we have that

$$\left|x_1^\top V^\top \phi\left(V(x_1+\varepsilon_1)\right)-x_1^\top V^\top \phi(Vx_1)\right| \leq L\left\|Vx_1\right\|_2\left\|V\varepsilon_1\right\|_2 = O\left(p\varepsilon\right), \tag{E.113}$$

with probability at least $1-2\exp\left(-c_1\log^2 d\right)$ due to Lemma E.5 and (E.70). For the second term in the RHS of (E.112), using the Hermite decomposition of ϕ, we have that

$$\left|x_1^\top V^\top \phi(Vx_1)-\mu_1 p\right| = \left|\sum_{k=1}^{p} v_k^\top x_1 \phi(v_k^\top x_1) - \mathbb{E}_{v_k}\left[v_k^\top x_1 \phi(v_k^\top x_1)\right]\right|. \tag{E.114}$$

Since ϕ is Lipschitz, we have that $\phi(v_k^\top x_1)$ is a sub-Gaussian random variable (with respect to v_k), and thus $v_k^\top x_1\phi(v_k^\top x_1)$ are p i.i.d. sub-exponential random variables. Thus, Bernstein inequality (see Theorem 2.8.1. in [Vershynin, 2018]) yields

$$\left|x_1^\top V^\top \phi(Vx_1)-\mu_1 p\right| = O\left(\sqrt{p}\log d\right) \tag{E.115}$$

with probability at least $1-2\exp\left(-c_3\log^2 d\right)$ over V. Then, plugging this and (E.113) in (E.112), together with the fact that $\mu_1 \neq 0$ by Assumption 4.2, we have

$$\left|x_1^\top V^\top \varphi(x_1+\varepsilon_1)-\mu_1 p\right| = O\left(p\varepsilon+\sqrt{p}\log d\right) = o(p), \tag{E.116}$$

with probability at least $1-2\exp\left(-c_4\log^2 d\right)$ over V. Considering now the last term in (E.110), we have

$$\left\|\varphi(x_1+\varepsilon_2)-\varphi(x_1+\varepsilon_1)\right\|_2 = O\left(\sqrt{p}\delta\right), \tag{E.117}$$

due to the Lipschitzness of ϕ and due to Lemma E.5. Thus, with probability $1-2\exp\left(-c_5\log^2 d\right)$, we have

$$\left|x_1^\top V^\top\left(\varphi(x_1+\varepsilon_2)-\varphi(x_1+\varepsilon_1)\right)\right| = O\left(p\delta\right), \tag{E.118}$$

where we used again Lemma E.5. Then, plugging (E.111), (E.116) and (E.118) in (E.110), an application of the triangle inequality yields

$$\left|a_1+a_2\right| p = O\left(\left|a_2\right| p\delta + \frac{p\log d}{\sqrt{d}}\right), \tag{E.119}$$

which gives the desired result. □

Lemma E.10. *Suppose $\varepsilon = o(1)$. Then, we have that*

$$|a_2| = O(\delta^{-1}), \qquad |a_1+a_2| = O(1), \tag{E.120}$$

with probability at least $1-2\exp\left(-c\log^2 d\right)$ over V and X.

Proof. The proof consists in considering the inner product of both sides of (E.65), namely

$$\begin{aligned}\varphi(x_2) = (a_1 + a_2)\varphi(x_1 + \varepsilon_1) \\ + a_2 \left((\varphi(x_1 + \varepsilon_2) - \varphi(x_1 + \varepsilon_1)) - \phi'(V(x_1 + \varepsilon_1)) \circ (V(\varepsilon_2 - \varepsilon_1))\right) \\ + a_2\, \phi'(V(x_1 + \varepsilon_1)) \circ (V(\varepsilon_2 - \varepsilon_1)),\end{aligned} \tag{E.121}$$

with $V(\varepsilon_2 - \varepsilon_1)$.

First, we have that the LHS reads

$$\left|(\varepsilon_2 - \varepsilon_1)^\top V^\top \varphi(x_2)\right| = O\left(\delta p\right), \tag{E.122}$$

due to Lemma E.5 with probability at least $1 - 2\exp\left(-c_1 \log^2 d\right)$ over V. Consider now $(\varepsilon_2 - \varepsilon_1)^\top V^\top \varphi(x_1 + \varepsilon_1)$. A net argument similar to the one used to obtain the third statement in Lemma E.1 yields

$$\left|(\varepsilon_2 - \varepsilon_1)^\top V^\top \varphi(x_1 + \varepsilon_1) - \mu_1 p \frac{(\varepsilon_2 - \varepsilon_1)^\top (x_1 + \varepsilon_1)}{d}\right| = O\left(\delta\sqrt{pd}\log d + \frac{\delta p}{d^2}\right), \tag{E.123}$$

with probability at least $1 - 2\exp\left(-c_2 \log^2 d\right)$ over V. This, together with (E.69) and Lemma E.9, gives

$$\left|(a_1 + a_2)(\varepsilon_2 - \varepsilon_1)^\top V^\top \varphi(x_1 + \varepsilon_1)\right| = O\left(\left(\left|a_2\right|\delta + \frac{\log d}{\sqrt{d}}\right)\left(\delta\sqrt{pd}\log d + \frac{\delta p}{d^2} + \delta\varepsilon p\right)\right), \tag{E.124}$$

with probability at least $1 - 2\exp\left(-c_3 \log^2 d\right)$ over V and X. Due to Lemma E.7, we have

$$\begin{aligned}&\left|(\varepsilon_2 - \varepsilon_1)^\top V^\top \left((\varphi(x_1 + \varepsilon_2) - \varphi(x_1 + \varepsilon_1)) - \phi'(V(x_1 + \varepsilon_1)) \circ (V(\varepsilon_2 - \varepsilon_1))\right)\right| \\ &\quad = O\left(p\delta^3 + d\log^2 d\delta^2\right) = o(p\delta^2),\end{aligned} \tag{E.125}$$

and due to Lemma E.8, we have

$$\left|(\varepsilon_2 - \varepsilon_1)^\top V^\top \left(\phi'(V(x_1 + \varepsilon_1)) \circ (V(\varepsilon_2 - \varepsilon_1))\right) - \frac{\left\|\varepsilon_2 - \varepsilon_1\right\|_2^2}{d} p\mu_1\right| = o(p\delta^2), \tag{E.126}$$

jointly with probability $1 - 2\exp\left(-c_4 \log^2 d\right)$ over V. Then, taking (E.122), (E.124), (E.125), and (E.126) together, we have that

$$\left|a_2\right| \frac{\left\|\varepsilon_2 - \varepsilon_1\right\|_2^2}{d} p\mu_1 = A_1 + A_2 + A_3 + A_4, \tag{E.127}$$

where

$$\begin{aligned}&|A_1| = O(\delta p), \\ &|A_2| = O\left(\left(\left|a_2\right|\delta + \frac{\log d}{\sqrt{d}}\right)\left(\delta\sqrt{pd}\log d + \frac{\delta p}{d^2} + \delta\varepsilon p\right)\right) = o(\delta p) + o\left(|a_2| p\delta^2\right), \\ &|A_3| = o\left(|a_2| p\delta^2\right), \\ &|A_4| = o\left(|a_2| p\delta^2\right).\end{aligned} \tag{E.128}$$

As $\mu_1 \neq 0$, this readily implies that

$$|a_2| = O\left(\delta^{-1}\right), \tag{E.129}$$

which, together with Lemma E.9 gives the desired result. □

Lemma E.11. *Suppose $\varepsilon = o(1)$. Then, we have that*

$$\left|a_2\tilde{\varphi}(x_2)^\top\left(\varphi(x_1+\varepsilon_2)-\varphi(x_1+\varepsilon_1)\right)\right| = o(p). \tag{E.130}$$

with probability at least $1-2\exp\left(-c\log^2 d\right)$ over V and X.

Proof. First, following the same decomposition considered in (E.82), we have

$$\begin{aligned}
&\left|a_2\tilde{\varphi}(x_2)^\top\left(\varphi(x_1+\varepsilon_2)-\varphi(x_1+\varepsilon_1)\right)\right| \\
&\leq \left|a_2\tilde{\varphi}(x_2)^\top\left(\left(\phi'(V(x_1+\varepsilon_1)+\zeta)-\phi'(V(x_1+\varepsilon_1))\right)\circ(V(\varepsilon_2-\varepsilon_1))\right)\right| \\
&\quad+\left|a_2\tilde{\varphi}(x_2)^\top\phi'(V(x_1+\varepsilon_1))\circ(V(\varepsilon_2-\varepsilon_1))\right|,
\end{aligned} \tag{E.131}$$

where each entry of ζ is such that $\zeta_k \in [0,[V(\varepsilon_2-\varepsilon_1)]_k]$ (or $\zeta_k \in [[V(\varepsilon_2-\varepsilon_1)]_k,0]$, if $[V(\varepsilon_2-\varepsilon_1)]_k$ is negative). The first term on the RHS can be bounded as

$$\begin{aligned}
&\left|a_2\tilde{\varphi}(x_2)^\top\left(\left(\phi'(V(x_1+\varepsilon_1)+\zeta)-\phi'(V(x_1+\varepsilon_1))\right)\circ(V(\varepsilon_2-\varepsilon_1))\right)\right| \\
&\quad\leq |a_2|\sum_{k=1}^{p}\left|\tilde{\phi}(v_k^\top x_2)\right|\min\left(2L,L'\left|v_k^\top(\varepsilon_2-\varepsilon_1)\right|\right)\left|v_k^\top(\varepsilon_2-\varepsilon_1)\right| \\
&\quad= |a_2|O\left(p\delta^2+d\log^2 d\delta\right)=O\left(p\delta+d\log^2 d\right),
\end{aligned} \tag{E.132}$$

with probability at least $1-2\exp\left(-c_1\log^2 d\right)$ over V and X, due to the same argument used in Lemma E.7 and where we used also Lemma E.10.

Next, let us look at the second term on the RHS of (E.131), and consider an $\epsilon\sqrt{d}$-net of $\sqrt{d}\,\mathbb{S}^{d-1}$ (as we did after (E.86)). Setting $\epsilon = d^{-2}$, we have that there exist $m^{(\xi)}$, $m^{(x_1)}$ and $m^{(x_2)}$ in $[M]$ such that $\left\|\xi - x^\epsilon_{m^{(\xi)}}\right\|_2 \leq 1/d^{3/2}$ (where ξ is defined in (E.84)), $\left\|(x_1+\varepsilon_1) - x^\epsilon_{m^{(x_1)}}\right\|_2 \leq 1/d^{3/2}$, $\left\|(x_2) - x^\epsilon_{m^{(x_2)}}\right\|_2 \leq 1/d^{3/2}$, and the size of the net is $M \leq e^{c_2 d\log d}$, where c_2 is an absolute positive constant. Consider three generic elements of this net: $x^\epsilon_{m^{(1)}}$, $x^\epsilon_{m^{(2)}}$, and $x^\epsilon_{m^{(3)}}$, and define

$$\begin{aligned}
T^{(m^{(1)},m^{(2)},m^{(3)})} = \Bigg|\sum_{k=1}^{p}\Big(&\tilde{\phi}(v_k^\top x^\epsilon_{m^{(1)}})\phi'\left(v_k^\top x^\epsilon_{m^{(2)}}\right)\left(v_k^\top x^\epsilon_{m^{(3)}}\right) \\
&-\mathbb{E}_{v_k}\left[\tilde{\phi}(v_k^\top x^\epsilon_{m^{(1)}})\phi'\left(v_k^\top x^\epsilon_{m^{(2)}}\right)\left(v_k^\top x^\epsilon_{m^{(3)}}\right)\right]\Big)\Bigg|.
\end{aligned} \tag{E.133}$$

Since $\left|\phi'\left(v_k^\top x^\epsilon_{m^{(1)}}\right)\right| \leq L$, each element of the sum is sub-exponential (in the probability space of v_k). Then, the argument based on Bernstein inequality used in Lemma E.8, after performing a union bound on the net, yields

$$\begin{aligned}
&T^{(m^{(x_2)},m^{(x_1)},m^{(\xi)})} = \\
&\left|\sum_{k=1}^{p}\left(\tilde{\phi}(v_k^\top x^\epsilon_{m^{(x_2)}})\phi'\left(v_k^\top x^\epsilon_{m^{(x_1)}}\right)\left(v_k^\top x^\epsilon_{m^{(\xi)}}\right)\right) - p\,\mathbb{E}\left[\tilde{\phi}\left(\rho^\epsilon_{x_2}\right)\phi'\left(\rho^\epsilon_{x_1}\right)\left(\rho^\epsilon_{\xi}\right)\right]\right| = o(p),
\end{aligned} \tag{E.134}$$

with probability at least $1-2\exp\left(-c_3 d\log^2 d\right)$ (due to Assumption 4.3), where we introduced $\rho^\epsilon_{x_1}$, $\rho^\epsilon_{x_2}$ and ρ^ϵ_{ξ} as three standard Gaussian variables with correlations

$$\begin{aligned}
\rho_{12} := \operatorname{corr}\left(\rho^\epsilon_{x_1},\rho^\epsilon_{x_2}\right) = \frac{(x^\epsilon_{m^{(x_1)}})^\top x^\epsilon_{m^{(x_2)}}}{d}, \qquad \rho_{1\xi} := \operatorname{corr}\left(\rho^\epsilon_{x_1},\rho^\epsilon_{\xi}\right) = \frac{(x^\epsilon_{m^{(x_1)}})^\top x^\epsilon_{m^{(\xi)}}}{d}, \\
\rho_{2\xi} := \operatorname{corr}\left(\rho^\epsilon_{x_2},\rho^\epsilon_{\xi}\right) = \frac{(x^\epsilon_{m^{(x_2)}})^\top x^\epsilon_{m^{(\xi)}}}{d}.
\end{aligned} \tag{E.135}$$

Note that

$$|\rho_{12}| = o(1), \qquad |\rho_{1\xi}| = o(1), \tag{E.136}$$

where the two bounds hold due to our net definition, due to $|x_1^\top x_2| = o(d)$ with probability at least $1 - 2\exp(-c_4 \log^2 d)$ (coming from Assumption 4.1), and due to (E.69). By Isserlis' theorem (or generalized Stein's lemma) we also have

$$\mathbb{E}\left[\tilde{\phi}(\rho_{x_2}^\epsilon)\phi'\left(\rho_{x_1}^\epsilon\right)\rho_\xi^\epsilon\right] = \rho_{2\xi}\mathbb{E}\left[\tilde{\phi}'(\rho_{x_2}^\epsilon)\phi'\left(\rho_{x_1}^\epsilon\right)\right] + \rho_{1\xi}\mathbb{E}\left[\tilde{\phi}(\rho_{x_2}^\epsilon)\phi''\left(\rho_{x_1}^\epsilon\right)\right]. \tag{E.137}$$

We have that

$$\left|\rho_{1\xi}\mathbb{E}\left[\tilde{\phi}(\rho_{x_2}^\epsilon)\phi''\left(\rho_{x_1}^\epsilon\right)\right]\right| \le |\rho_{1\xi}|\,\mathbb{E}\left[\tilde{\phi}(\rho_{x_2}^\epsilon)^2\right]^{1/2}\mathbb{E}\left[\phi''\left(\rho_{x_1}^\epsilon\right)^2\right]^{1/2} = o(1), \tag{E.138}$$

where we used (E.136), $|\phi''| \le L'$ (L' being the Lipschitz constant of ϕ') and that $\tilde{\phi}$ is Lipschitz due to Assumption 4.2. To study the first term on the LHS of (E.137), notice that the Hermite coefficient of order 0 of $\tilde{\phi}'$ is 0, since the Hermite coefficient of order 1 of $\tilde{\phi}$ is 0 by definition. Thus, denoting by $\mu_r', \tilde{\mu}_r'$ the r-th Hermite coefficient respectively of $\phi', \tilde{\phi}'$, we have

$$\begin{aligned}
\left|\rho_{2\varepsilon}\mathbb{E}\left[\tilde{\phi}'(\rho_{x_2}^\epsilon)\phi'\left(\rho_{x_1}^\epsilon\right)\right]\right| &\le |\rho_{2\varepsilon}|\left|\sum_{r=1}^\infty \tilde{\mu}_r'\mu_r'\rho_{12}^r\right| \le |\rho_{12}|\sum_{r=1}^\infty |\tilde{\mu}_r'\mu_r'| \\
&\le |\rho_{12}|\sqrt{\sum_{r=1}^\infty (\tilde{\mu}_r')^2}\sqrt{\sum_{r=1}^\infty (\mu_r')^2} \\
&\le |\rho_{12}|\,\mathbb{E}\left[\tilde{\phi}'(\rho_{x_2}^\epsilon)^2\right]^{1/2}\mathbb{E}\left[\phi'\left(\rho_{x_1}^\epsilon\right)^2\right]^{1/2} = o(1),
\end{aligned} \tag{E.139}$$

where in the last step we used (E.136), $|\phi'| \le L$ and that $\tilde{\phi}$ is Lipschitz due to Assumption 4.2. Putting (E.138) and (E.139) in (E.137), and plugging this in (E.134) yields

$$\left|\sum_{k=1}^p \tilde{\phi}(v_k^\top x_{m^{(x_2)}}^\epsilon)\phi'\left(v_k^\top x_{m^{(x_1)}}^\epsilon\right)\left(v_k^\top x_{m^{(\xi)}}^\epsilon\right)\right| = o(p), \tag{E.140}$$

with probability at least $1 - 2\exp(-c_5 \log^2 d)$ over V and X.

Then, following a similar argument as the one that led to (E.108), we obtain

$$\left|\sum_{k=1}^p \tilde{\phi}(v_k^\top x_{m^{(x_2)}}^\epsilon)\phi'\left(v_k^\top x_{m^{(x_1)}}^\epsilon\right)\left(v_k^\top x_{m^{(\xi)}}^\epsilon\right) - \tilde{\phi}(v_k^\top x_2)\phi'\left(v_k^\top(x_1+\varepsilon_1)\right)\left(v_k^\top \xi\right)\right| = o(p), \tag{E.141}$$

with probability at least $1 - 2\exp\left(-c_6 \log^2 d\right)$ over V due to Lemma E.5. Putting (E.140) and (E.141) together with the result of Lemma E.10 yields

$$\left|a_2\tilde{\varphi}(x_2)^\top \phi'(V(x_1+\varepsilon_1)) \circ (V(\varepsilon_2 - \varepsilon_1))\right| = o(p), \tag{E.142}$$

with probability at least $1 - 2\exp\left(-c_7 \log^2 d\right)$ over V and X. This last equation, when plugged in (E.131) together with (E.132), gives the desired result. □

Proof of Theorem 4.5. Let us consider

$$\varphi(x_2) = a_1\varphi(x_1 + \varepsilon_1) + a_2\varphi(x_1 + \varepsilon_2), \tag{E.143}$$

where

$$\|x_1 + \varepsilon_1\|_2 = \|x_1 + \varepsilon_2\|_2 = \sqrt{d}. \tag{E.144}$$

Suppose by contradiction that there exists a solution to the equations above such that $\varepsilon = o(1)$. Take the inner product of both sides of (E.143) with the vector $\tilde{\varphi}(x_2)$, namely

$$\tilde{\varphi}(x_2)^\top \varphi(x_2) = (a_1 + a_2)\tilde{\varphi}(x_2)^\top \varphi(x_1 + \varepsilon_1) + a_2\tilde{\varphi}(x_2)^\top \left(\varphi(x_1 + \varepsilon_2) - \varphi(x_1 + \varepsilon_1)\right). \tag{E.145}$$

For the LHS, we have that

$$\tilde{\varphi}(x_2)^\top \varphi(x_2) = \Theta(p), \tag{E.146}$$

with probability at least $1 - 2\exp\left(-c_1 \log^2 d\right)$, due to the first statement of Lemma E.1. For the first term of the RHS, we have

$$\begin{aligned} &\left|(a_1 + a_2)\,\tilde{\varphi}(x_2)^\top \varphi(x_1 + \varepsilon_1) - (a_1 + a_2)\,\tilde{\varphi}(x_2)^\top \varphi(x_1)\right| \\ &\qquad \le |a_1 + a_2|\,\|\tilde{\varphi}(x_2)\|_2\, L\, \|V\|_{\mathsf{op}}\, \|\varepsilon_1\|_2 = o(p), \end{aligned} \tag{E.147}$$

with probability at least $1 - 2\exp\left(-c_2 \log^2 d\right)$, due to Lemmas E.5 and E.10. Then, using the Hermite expansion of $\tilde{\phi}$ and ϕ we have

$$\left|\tilde{\varphi}(x_2)^\top \varphi(x_1) - p\sum_{r=3}^{\infty}\left(\frac{x_2^\top x_1}{d}\right)^r\right| = o(p), \tag{E.148}$$

with probability at least $1 - 2\exp\left(-c_3 \log^2 d\right)$ due to Bernstein inequality. As we have $|x_2^\top x_1| = o(d)$ with this same probability due to Assumption 4.1, we readily obtain

$$\left|(a_1 + a_2)\tilde{\varphi}(x_2)^\top \varphi(x_1)\right| = o(p), \tag{E.149}$$

where we also used the bound on $|a_1 + a_2|$ from Lemma E.10.

Finally, due to Lemma E.11, we have

$$\left|a_2\tilde{\varphi}(x_2)^\top \left(\varphi(x_1 + \varepsilon_2) - \varphi(x_1 + \varepsilon_1)\right)\right| = o(p) \tag{E.150}$$

with probability at least $1 - 2\exp\left(-c_4 \log^2 d\right)$.

Plugging (E.146), (E.147), (E.149) and (E.150) in (E.145) generates a contradiction with high probability. This implies that, with probability at least $1 - 2\exp\left(-c_5 \log^2 d\right)$, we have $\varepsilon = \Omega(1)$. Taking the intersection between this event and the one described by Theorem 4.4, we obtain the thesis. □

E.3 Implementation details

Computational resources. We executed all the experiments on a machine equipped with two GPUs (an NVIDIA GeForce RTX 3090, 24 GB VRAM and an NVIDIA RTX A5000, 24 GB VRAM), an Intel(R) Core(TM) i9-10920X CPU @ 3.50GHz and 128 GB of RAM.

Training procedure of neural networks. By default, we train both two-layer and deep residual networks with full-batch gradient descent with step size 10^{-5} on square loss for 10^6 steps, unless stated otherwise. For the classification experiments on random features and two-layer neural networks, we set the step size to 10^{-3} and optimize logistic loss (in the case of binary classification) or cross-entropy loss (in the case of multi-class classification) with full-batch gradient descent for 10^6 steps.

Optimization details for the reconstruction algorithm. We provide precise implementation details of the reconstruction algorithm. The inputs of our procedure are the model's optimal weights θ^* and the structure of f_{RF}, *i.e.* , V and ϕ. As already mentioned in Section 4.5.2, in the case of neural networks where initialization is not zero, we define $\theta^* \triangleq \theta_*^{(L)} - \theta_0^{(L)}$, where θ_0 are the initial parameters and θ^* are the trained parameters after gradient descent converged on the training set. The superscript $^{(L)}$ indicates the weights of the last layer. For neural networks we also assume access to their trained weights of all layers $\{\theta^{*(1)}, \dots, \theta^{*(L)}\}$, but we only require knowledge of the initialization of the last layer $\theta_0^{(L)}$ and their internal structure. The reconstruction problem is solved by minimizing the following objective via gradient descent with momentum:

$$\hat{X}^\star = \underset{\hat{X}:\|\hat{x}_i\|_2=\sqrt{d}}{\arg\min} \left\| P_{\hat{\Phi}}^{\perp} \theta^* \right\|_2^2 . \tag{E.151}$$

We initialize the rows $\hat{x}_i$ as i.i.d. standard Gaussian vectors. To efficiently compute the objective in (E.151) at each gradient descent iterate, we leverage the fact that, by definition, $P_{\hat{\Phi}}^{\perp}$ is symmetric ($(P_{\hat{\Phi}}^{\perp})^\top = P_{\hat{\Phi}}^{\perp}$) and idempotent ($(P_{\hat{\Phi}}^{\perp})^2 = P_{\hat{\Phi}}^{\perp}$). Expanding according to the definitions and properties above,

$$\begin{aligned}
\mathcal{L}(\hat{X}) = \left\| P_{\hat{\Phi}}^{\perp} \theta^* \right\|_2^2 &= [P_{\hat{\Phi}}^{\perp} \theta^*]^\top P_{\hat{\Phi}}^{\perp} \theta^* \\
&= [\theta^*]^\top [P_{\hat{\Phi}}^{\perp}]^\top P_{\hat{\Phi}}^{\perp} \theta^* \\
&= [\theta^*]^\top P_{\hat{\Phi}}^{\perp} \theta^* \\
&= [\theta^*]^\top (I - \hat{\Phi}^+ \hat{\Phi}) \theta^* \\
&= [\theta^*]^\top \theta^* - [\theta^*]^\top \hat{\Phi}^\top (\hat{\Phi}\hat{\Phi}^\top)^{-1} \hat{\Phi} \theta^* \\
&= [\theta^*]^\top \theta^* - [\theta^*]^\top \hat{\Phi}^\top \alpha \ .
\end{aligned}$$

Here, α is the solution of the system of n linear equations in n unknowns $(\hat{\Phi}\hat{\Phi}^\top)\, \alpha = \hat{\Phi}\theta^*$ which can be numerically computed via the conjugate gradient method [Hestenes et al., 1952], skipping the need of inverting explicitly $\hat{\Phi}\hat{\Phi}^\top \in \mathbb{R}^{n \times n}$ at each iteration. After the gradient descent update on $\hat{X}$, we then normalize its rows $\hat{x}_i$ to force them to lie on the d-dimensional sphere $\sqrt{d}\, \mathbb{S}^{d-1}$. More precisely, $\hat{x}_i \leftarrow \sqrt{d} \cdot \hat{x}_i / \|\hat{x}_i\|_2$, thus respecting the fact that minimization should happen on the sphere as in (E.151). Unless stated otherwise, the step size is set by default to $2 \cdot 10^3$ for the CIFAR-10 experiments and to 20 for the experiments on synthetic data. On Tiny-ImageNet we use a step size of $8 \cdot 10^3$. Momentum is set to 0.9. We consider the reconstruction optimization converged when the normalized reconstruction loss $\mathcal{L}(\hat{X})/\|\theta^*\|_2^2$ drops below 10^{-7}. Normalizing by $\|\theta^*\|_2^2$ makes the loss of order 1 at the beginning of the optimization so that the chosen threshold of 10^{-7} corresponds to the effective numerical resolution in floating point representation. Unless explicitly mentioned otherwise, in every experiment we perform, the optimization is run until the reconstruction loss has converged.

Statistical robustness. To ensure that our findings are not tied to a particular random initialization, all experiments are repeated over multiple random seeds. Each seed induces independent initializations of network parameters and of the reconstruction variables $\hat{X}$. All the quantitative results reported in the chapter correspond to averages (and variability) across these runs.

CIFAR-10 protocol. For random features experiments, we construct a balanced subset from the CIFAR-10 [Krizhevsky et al., 2009] training split by iterating once through the data and collecting the first $n/2$ occurrences of the "*frog*" class (negative) and $n/2$ of the "*truck*" class (positive), yielding n samples. Each image is flattened to $d = 3 \cdot 32 \cdot 32 = 3072$ and concatenated into $X \in \mathbb{R}^{n \times d}$. We standardize using subset statistics computed over the selected n examples. Targets are $Y \in \{\pm 1\}^n$, with the first $n/2$ entries set to -1 and the remaining $n/2$ to $+1$. We use "*frog*" vs. "*truck*" purely as a convenient benchmark. Our conclusions do not hinge on category semantics, and we observed the same qualitative behavior across alternative class pairs. For multiclass experiments, we use the same balanced construction; targets are one-hot encoded as 10-dimensional vectors with a single 1 at the true class index (and 0 elsewhere). Also for ResNet and vision transformer runs we use the same protocol as for random features, but images are kept in tensor form ($\mathbb{R}^{3\times32\times32}$) and normalized per channel using the full CIFAR-10 training split statistics (mean $[0.4914, 0.4822, 0.4465]$, std $[0.247, 0.243, 0.261]$).

Tiny-ImageNet protocol. We construct a balanced subset from the Tiny-ImageNet [Le and Yang, 2015] training split by iterating once through the data and collecting the first $n/2$ occurrences of *class 2* (negative) and $n/2$ of *class 116* (positive), yielding n samples. Each image is flattened to $d = 3 \cdot 64 \cdot 64 = 12288$ and concatenated into $X \in \mathbb{R}^{n \times d}$. We standardize using subset statistics computed over the selected n samples. Targets are $Y \in \{\pm 1\}^n$, with the first $n/2$ entries set to -1 and the remaining $n/2$ to $+1$. We have randomly drawn *class 2* vs. *class 116* and used this pair purely as a convenient benchmark. Also in this case, our conclusions do not hinge on category semantics.

E.3.1 Additional details on deep residual networks

In this section, we give precise implementation details of the class of deep residual networks we consider. We focus on ResNet-style architectures [He et al., 2016] adapted to CIFAR-10 images while preserving the canonical residual topology. Specifically, we remove batch normalization and max pooling layers so that feature map resolution and statistics are maintained throughout. Apart from this, the residual block structure and overall layout are preserved. More formally, let $\phi(\cdot) = \mathrm{ReLU}(\cdot)$ applied element-wise and let $C^{k\times k}_{a\to b}[\cdot]$ denote a bias-free 2D convolution with kernel size $k \times k$, stride 1 and padding 1, mapping from a to b channels. Let a residual block be

$$B(u) = \phi(u + C^{3\times3}_{h\to h}[\phi(C^{3\times3}_{h\to h}[u])]) \ . \tag{E.152}$$

Given an image $x \in \mathbb{R}^{3\times32\times32}$ (as in the case of CIFAR-10 examples), we consider in our experiments deep residual networks with architecture

$$z_0 = C^{28\times28}_{3\to h}[x] \qquad z_b = B(z_{b-1}) \quad \text{for } b = 1, \dots, 4 \ , \qquad f_{\mathsf{RN}}(x) = [\theta^{(L)}]^\top \mathrm{vec}(z_4) \ , \tag{E.153}$$

with $\theta^{(L)} \in \mathbb{R}^{h\cdot7\cdot7}$ the last layer's parameters. We denote with $\mathrm{vec}(\cdot)$ the flattening of the dimensions of the 3D tensor z_4 to a vector. The choice of dimensionality for $\theta^{(L)}$ is motivated by the first convolution mapping the spatial size for CIFAR-10 images to $h \times 7 \times 7$ and the blocks preserving it. We select the number of output channels h based on the choice of the

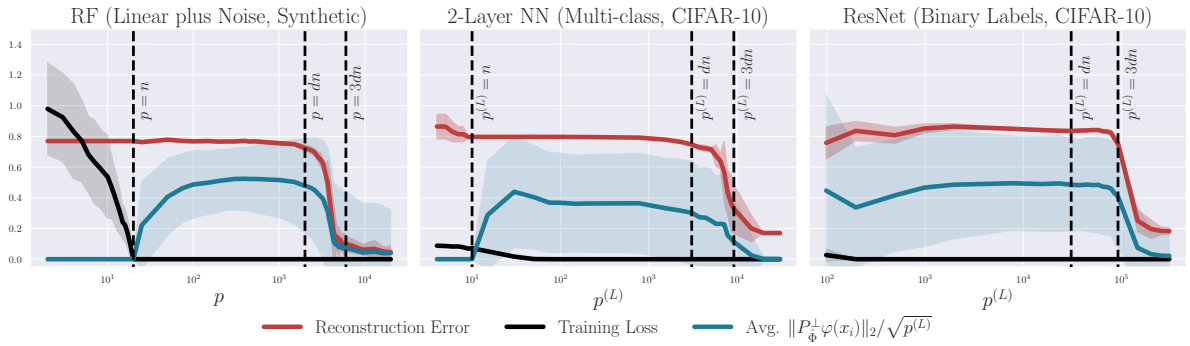


Figure E.1: **Features of the training dataset Φ are spanned by the features of the reconstructed dataset $\hat{\Phi}$.** We consider RF regression with ReLU activation on i.i.d. samples uniformly distributed on the d-dimensional sphere ($d = 100, n = 20$), 10-class one-hot regression with a 2-layer ReLU network trained with gradient descent on CIFAR-10 ($n = 10$) and regression on binary labels (*frogs vs. trucks*) with a ResNet trained with gradient descent on CIFAR-10 ($n = 10$). We report mean (solid line) and standard deviation (shaded area) for the reconstruction error (in red), for the training loss (in black) and for the per-feature orthogonal residual of projecting Φ onto $\mathrm{Span}\{\mathrm{rows}(\hat{\Phi})\}$ (in blue), as the number of parameters increases. We indicate with $p^{(L)}$ the number of parameters in the last layer. Statistics are computed across 10 distinct random seeds.

number of parameters in the last layer $p^{(L)}$ as these two quantities are tied together, being $h = p^{(L)}/7^2$, eventually truncated to the nearest integer. Parameters of the first convolutional layer are initialized as $\theta_{i,j}^{(0)} \sim \mathcal{N}(0, 1/d)$ i.i.d. , while for $l > 1$ parameters are initialized as $\theta_{i,j}^{(l)} \sim \mathcal{N}(0, 1/\mathrm{fan\_in})$ i.i.d. . Here, $\mathrm{fan\_in}$ equals the number of input units (for convolutional layers: $h \cdot k^2$; for the last linear layer: $p^{(L)}$).

E.4 Additional numerical results

E.4.1 Further results on span inclusion

As mentioned at the end of Section 4.4 and in the caption of Figure 4.2, we test whether the features computed on the training dataset are spanned by the features of the reconstructed examples. To this end, we calculate the average per-feature orthogonal residual, formally defined as

$$r(\hat{\Phi}) = \frac{1}{n\sqrt{p}} \sum_{i=1}^{n} \left\| P_{\hat{\Phi}}^{\perp} \varphi(x_i) \right\|_2 , \tag{E.154}$$

where $P_{\hat{\Phi}}^{\perp} = I - P_{\hat{\Phi}}$ projects onto the orthogonal complement of $\mathrm{Span}\{\mathrm{rows}(\hat{\Phi})\}$. Thus $r(\hat{\Phi}) = 0$ if and only if every $\varphi(x_i)$ lies in $\mathrm{Span}\{\mathrm{rows}(\hat{\Phi})\}$. The normalization by $n\sqrt{p}$ makes $r(\hat{\Phi})$ of order 1 so that values of $r(\hat{\Phi}) \ll 1$ indicate numerically negligible residuals (*i.e.* , effective span inclusion).

Figure E.1 summarizes the interpolation-reconstruction behavior across RF regression (synthetic data, same setting of Figure 4.3), two-layer ReLU neural networks (CIFAR-10, multi-class, same setting of left Figure 4.6), and deep residual networks (CIFAR-10, binary, same setting of right Figure 4.6). Both the reconstruction error (red) and the per-feature orthogonal residual (blue) remain large until the model crosses the threshold $p \approx dn$ ($p^{(L)} \approx dn$ for neural networks), after which they drop sharply. This trend is consistent with Figure 4.2: successful reconstruction occurs precisely when $\varphi(x_i) \in \mathrm{Span}\{\mathrm{rows}(\hat{\Phi})\}$ for all i, *i.e.* , when

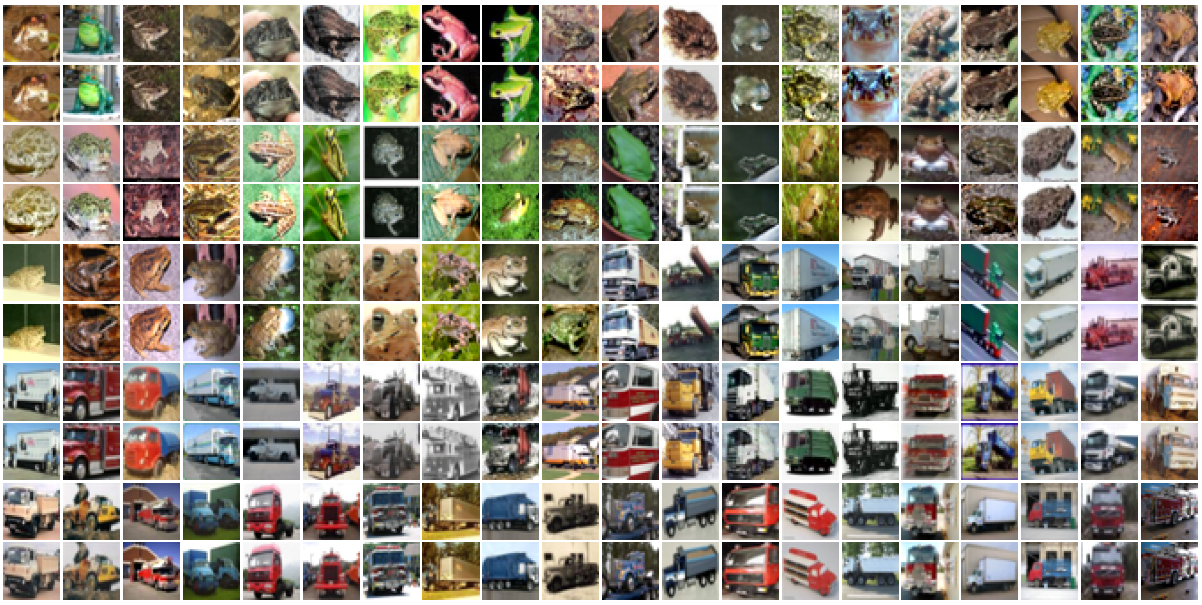


Figure E.2: **Reconstructing CIFAR-10 training images from an RF model with the activation function satisfying Assumption 4.2.** We repeat Figure 4.4 with the same seed, reconstructing CIFAR-10 training images ($n = 100$) from an RF model with activation function $\phi(z) = \mathrm{ReLU}(z) + \tanh(z)$ on binary labels at $p = 10dn$. Odd rows report the ground truth images, while even rows the reconstructed ones. In this experiment the activation function is consistent with Assumption 4.2, and indeed, this excludes the possibility of reconstructing sign-flipped versions of the originals.

$\|P_{\hat{\Phi}}^{\perp}\varphi(x_i)\|_2 \approx 0$. Also in these settings, for $p \leq n$, the optimization converges to the non-degenerate case $P_{\hat{\Phi}} = I$, which makes the orthogonal residual trivially zero.

E.4.2 Data reconstruction on RF with mixed-parity activations

Figure E.2 repeats the setup of Figure 4.4 on CIFAR-10 (binary labels, $n = 100$) using the *same random seed* and optimization protocol, but replacing ReLU with the activation $\phi(z) = \mathrm{ReLU}(z) + \tanh(z)$ that has high-order Hermite coefficients of different parities. In contrast to Figure 4.4, no sign-flipped reconstructions appear. This outcome is precisely what Assumption 4.2 and Remark 4.4 predict: ReLU alone violates the assumption (its odd Hermite coefficients $\mu_{2\ell+1}$ vanish for $\ell > 1$), allowing x_i and $-x_i$ to be indistinguishable under the reconstruction loss. Adding $\tanh$ supplies non-zero coefficients of opposite parity, restoring identifiability of the sign and removing this degeneracy.

E.4.3 Additional ablation studies

Effects of different learning rates in the reconstruction optimization. To assess the effect of the step size on the success of the data reconstruction, we have conducted an additional ablation using Binary CIFAR-10 ($n = 20$) on the RF model. Starting from the base step size $\eta^* = 2 \cdot 10^3$, we have considered $\eta \in \{\eta^*/4, \eta^*/2, \eta^*, 2\eta^*, 4\eta^*\}$ across 10 distinct random seeds. In Figure E.3, we plot the reconstruction error, the total number of iterations for which the reconstruction algorithm is run, and the final reconstruction loss $\mathcal{L}$, as a function of the number of parameters p. For larger step sizes ($2\eta^*$ and $4\eta^*$), we observe oscillatory behavior of the reconstruction loss that prevents convergence within a reasonable budget. Guided by the smallest step size ($\eta^*/4$), for which the loss consistently converges to zero within at most $6 \cdot 10^4$ iterations across seeds and number of parameters p, we cap the optimization at $6 \cdot 10^4$ iterations in this experiment. As shown in Figure E.3, for all step sizes

that converge within this budget ($\eta^*/4, \eta^*/2, \eta^*$), the reconstruction error drops to zero once $p \gg dn$, indicating that the reconstruction procedure is robust to the choice of the step size provided it is not taken excessively large.

Number of reconstructed samples different from number of training samples. So far, we only discussed the results of optimizing $\mathcal{L}(\hat{X})$ when setting $\hat{X}$ to be a matrix in $\mathbb{R}^{n \times d}$. In Figure E.4, we numerically investigate the effects of setting $\hat{X}$ to be a matrix of size $\hat{n} \times d$, when $\hat{n} \neq n$. Specifically, we consider Binary CIFAR-10 with $n = 50$ (25 "frog" images vs. 25 "truck" images) and a RF model with $p = 10dn$. When reconstructing fewer samples ($\hat{n} < n$), we observe that the outputs of the reconstruction often mix the structure coming from multiple ground-truth images. Intuitively, this translates in the fact that the reconstruction seems to be minimized when $\hat{X}$ has rows that are a superposition of multiple training data, rather than a subset of the training data themselves. Increasing $\hat{n}$ progressively reduces this noise. Conversely, when reconstructing with more rows than training samples ($\hat{n} > n$), we find that 50 of the recovered images match the $n = 50$ training samples, while the extra $\hat{n} - n = 10$ reconstructions are simply duplicates. From a practical perspective, this means that, in order to estimate n, it suffices to iteratively increase $\hat{n}$ until the first duplicate appears.

Reconstruction from vision transformer architecture. In Figure E.5, we provide numerical results on vision transformers trained on CIFAR-10 in the multi-class setting, using the same reconstruction procedure. We study randomly initialized vision transformers [Dosovitskiy et al., 2021] akin to ViT-B/16 on which we vary the embedding dimension trained on one-hot encoded labels from the first five classes of the CIFAR-10 dataset ($n = 5$). Vision transformers display a similar phenomenology to fully connected and residual architectures: when the number of parameters $p^{(L)}$ exceeds dn, the reconstruction error goes down and images are recovered successfully.

Effects of weight regularization. So far, we discussed reconstruction from the weights of a trained model without regularization, as expressed in Eq. (4.2). Adding a ridge parameter $\lambda > 0$ would change the training loss as

$$\mathcal{L}_{train}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left(\varphi(x_i)^\top \theta - y_i \right)^2 + \lambda \|\theta\|_2^2, \tag{E.155}$$

whose unique minimizer is

$$\theta^* = \Phi^\top \left(\Phi \Phi^\top + n\lambda I \right)^{-1} Y. \tag{E.156}$$

For a fixed value of λ (not dependent on d, n and p), and due to Eq. (E.1), we have

$$\|n\lambda I\|_{\mathsf{op}} = O(n), \qquad \lambda_{\min} \left(\Phi \Phi^\top \right) = \Omega(p). \tag{E.157}$$

Intuitively this suggests that, as p grows large, the effect of a fixed regularization term λ becomes negligible, and reconstruction is still possible. This is verified numerically in Figure E.7, where we see that even for very large values of λ we find a qualitatively similar threshold for successful reconstruction.

Effects of pruning. In Figure E.8, we explore numerically the behavior of our reconstruction algorithm on a pruned network, following the same synthetic setup of Figure 4.3. We first obtain $\theta^* = \Phi^+ Y$ and then run Optimal Brain Surgeon [Hassibi and Stork, 1992], which in our context (convex problem on square loss) provably incurs in the minimal increase in training loss, given a fixed sparsity ratio. We select either $d = 30, n = 20$ (left plot of Figure 4.3) or $d = 100, n = 100$ (right plot of Figure 4.3), reporting the results for several sparsity ratios. As expected, pruning seems to increase the number of parameters needed for data reconstruction.

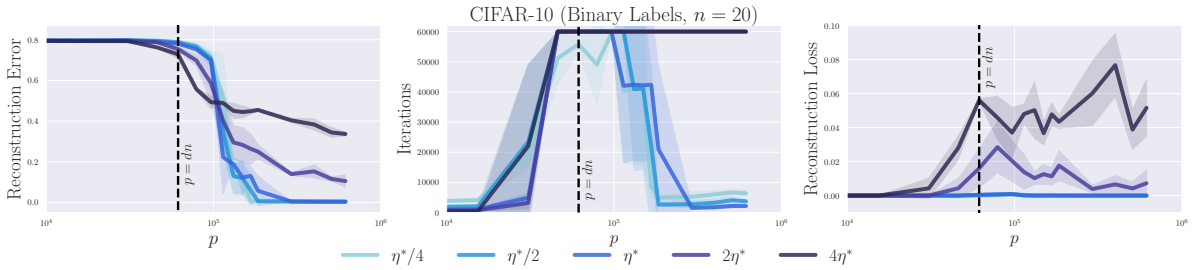


Figure E.3: **Ablation on step size used for the reconstruction procedure.** We consider RF regression with ReLU activation on binary labels (*frogs vs. trucks*) with $n = 20$ images (10 examples per class) from CIFAR-10. Starting from the base step size $\eta^* = 2 \cdot 10^3$ used in all the CIFAR-10 experiments, we report the reconstruction error (left), total number of training iterations needed by the reconstruction algorithm (center) and reconstruction loss (right) as a function of the number of parameters p, aggregated over 10 distinct random seeds. Guided by the smallest step size ($\eta^*/4$), for which the loss consistently converges to zero within at most $6 \cdot 10^4$ iterations across seeds and number of parameters p, we cap the optimization at $6 \cdot 10^4$ iterations.

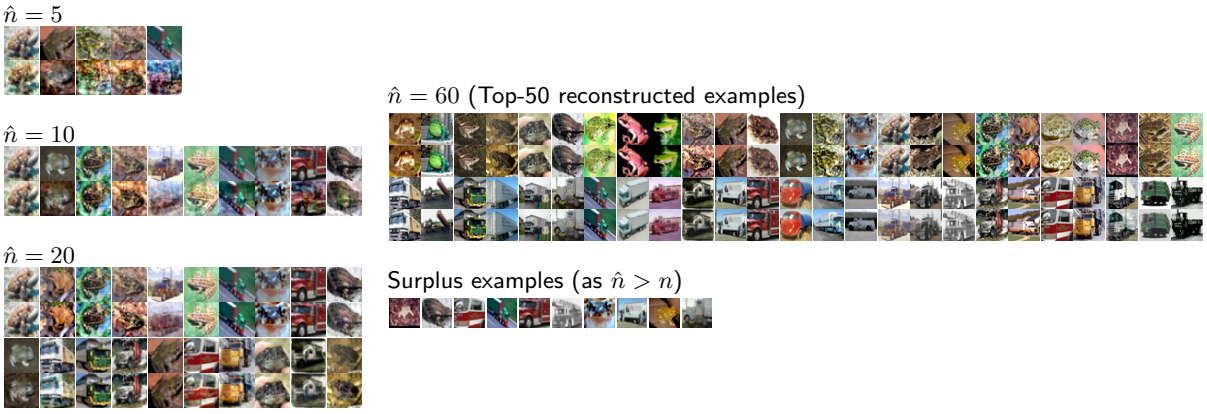


Figure E.4: **Dataset reconstruction without knowing the exact number of training samples n.** We fit an RF model with ReLU activation and $p = 10dn$ on binary labels (*frogs vs. trucks*) with $n = 50$ images (25 examples per class) from CIFAR-10. We then optimize the reconstruction loss $\mathcal{L}(\hat{X})$ with $\hat{X} \in \mathbb{R}^{\hat{n} \times d}$ and the number of reconstructed samples $\hat{n}$ different from n. We assess both the case when reconstructing fewer samples ($\hat{n} < n$, left column) and more samples ($\hat{n} > n$, right column) than the ground-truth number of samples n. In the latter case, we also plot the extra $\hat{n} - n = 10$ examples.

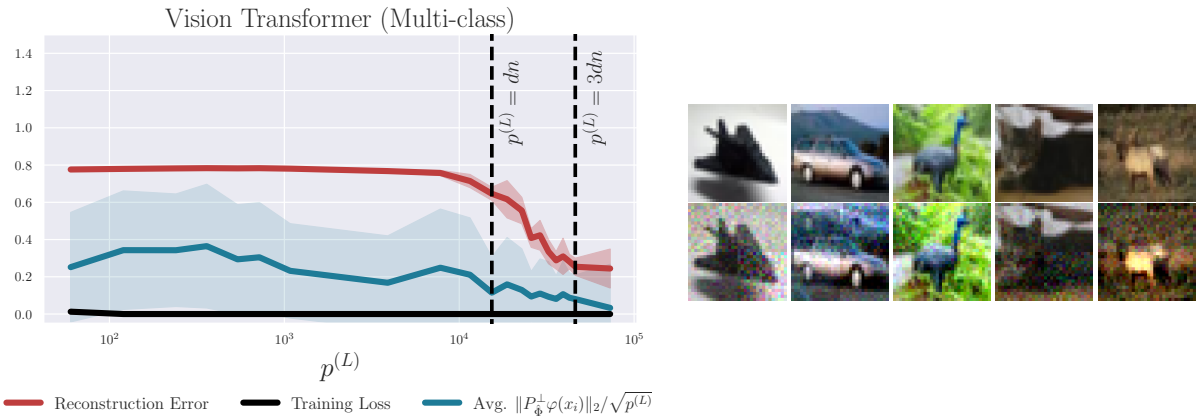


Figure E.5: **Thresholds for label fitting and data reconstruction for Vision Transformers trained with gradient descent on $n = 5$ CIFAR-10 images.** (Left) We consider regression with the square loss, training Vision Transformers on the one-hot encoding of 5-class labels. We report mean (solid line) and standard deviation (shaded area) for the reconstruction error (in red), for the training loss (in black) and for the per-feature orthogonal residual of projecting Φ onto $\mathrm{Span}\{\mathrm{rows}(\hat{\Phi})\}$ (in blue), as the number of parameters in the last layer $p^{(L)}$ increases. Statistics are computed across 10 distinct random seeds. (Right) Results of the reconstruction when $p^{(L)} = 4dn$. The first row reports the ground truth images, while the second row the reconstructed ones.

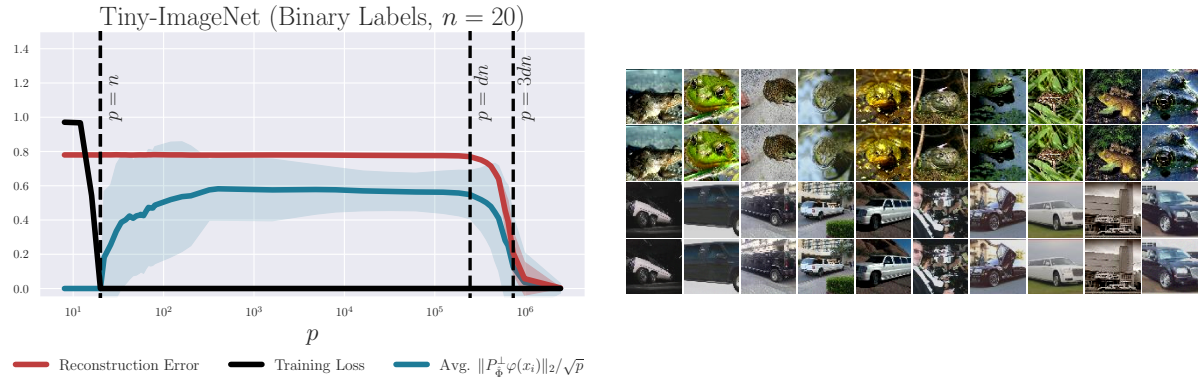


Figure E.6: **Thresholds for label fitting and data reconstruction on Tiny-ImageNet.** (Left) We consider RF regression with ReLU activation on binary labels (*class 2 vs. class 116*) with $n = 20$ images (10 examples per class) from Tiny-ImageNet ($d = 12288$). We report mean (solid line) and standard deviation (shaded area) for the reconstruction error (in red), for the training loss (in black) and for the per-feature orthogonal residual of projecting Φ onto $\mathrm{Span}\{\mathrm{rows}(\hat{\Phi})\}$ (in blue), as the number of parameters p increases. Statistics are computed across 10 distinct random seeds. (Right) Results of the reconstruction when $p = 10dn$. Odd rows report the ground truth images, while even rows the reconstructed ones which are all visually very similar.

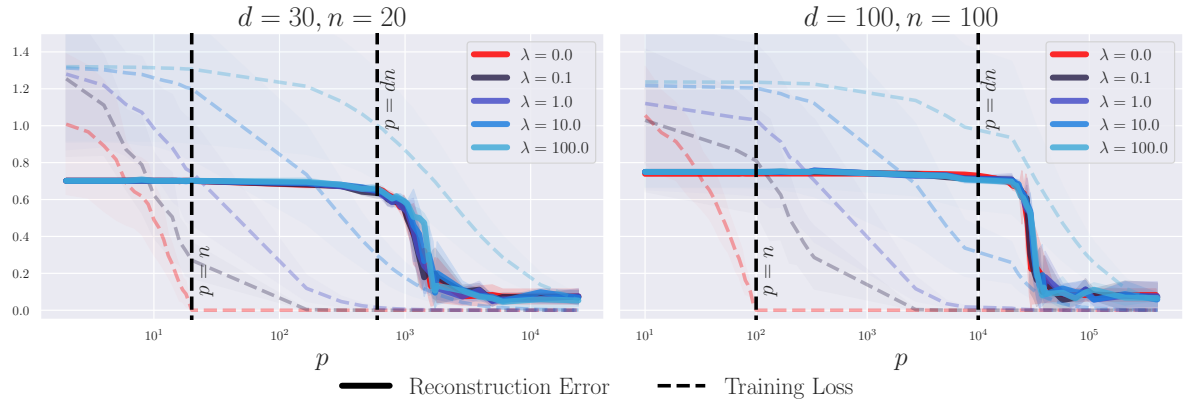


Figure E.7: **Thresholds for label fitting and data reconstruction in the presence of regularization.** We consider RF regression with ReLU activation, fitting a noisy linear model with ℓ_2 penalty (ridge). The training data is i.i.d. uniformly drawn from the d-dimensional sphere. We report mean (solid/dashed lines) and standard deviation (shaded areas) for both the reconstruction error (solid lines) and training loss (dashed lines) as the number of parameters p increases, at different choices of input dimensions d and number of samples n. Statistics are computed across 10 distinct random seeds. Lighter colors refer to stronger regularization, which does not significantly deviate from the behavior of training without regularization (red lines, occluded).

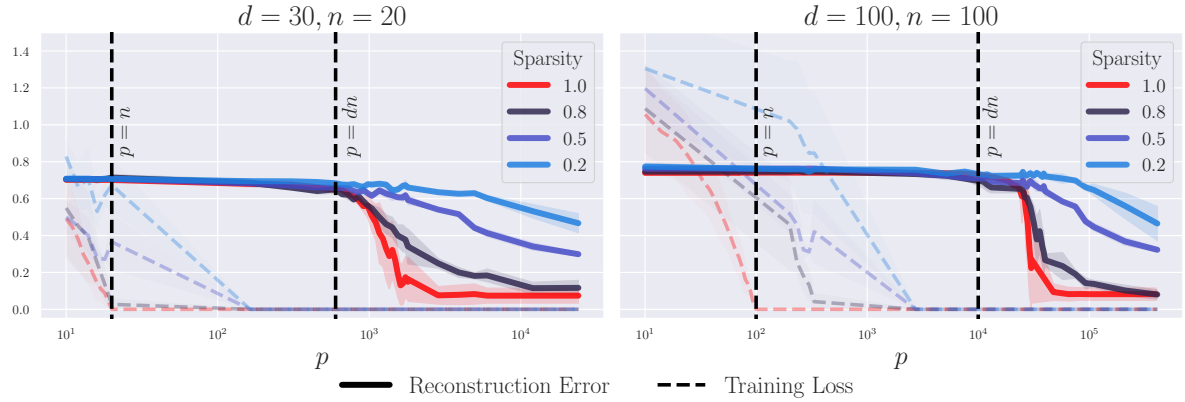


Figure E.8: **Thresholds for label fitting and data reconstruction of pruned models.** We consider RF regression with ReLU activation, fitting a noisy linear model. The training data is i.i.d. uniformly drawn from the d-dimensional sphere. After fitting, we prune the trained parameters θ^* with Optimal Brain Surgeon [Hassibi and Stork, 1992] to a desired sparsity ratio (*i.e.* , p'/p with p' the number of remaining parameters). We report mean (solid/dashed lines) and standard deviation (shaded areas) for both the reconstruction error (solid lines) and training loss (dashed lines) as the number of parameter p increases, at different choices of input dimensions d and number of samples n. Statistics are computed across 10 distinct random seeds. Lighter colors refer to lower sparsity ratios, compared against the unpruned baselines (red lines).

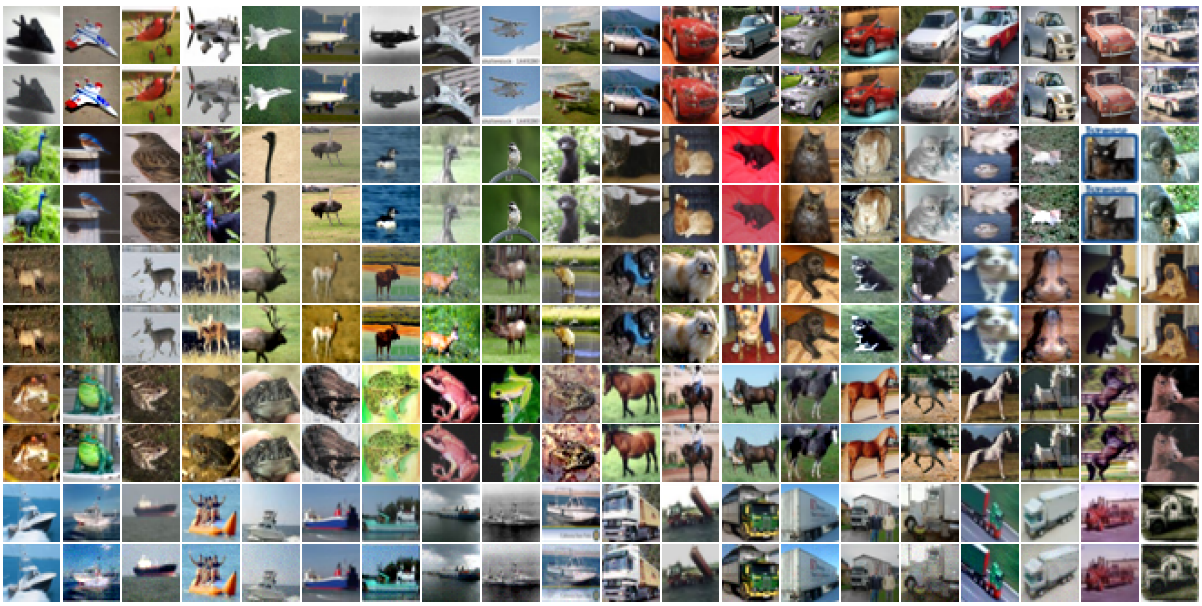


Figure E.9: **Multi-class training data reconstructed from a neural network trained with cross-entropy on CIFAR-10.** We repeat the experiment of Figure 4.7 by training a two-layer ReLU network on $n = 100$ images from CIFAR-10 dataset (10 examples per class) with gradient descent. In this experiment, we cast the training procedure as multi-class classification on *cross-entropy loss*. The number of parameters in the last layer of the network is $p^{(L)} = 4dn$.

APPENDIX F

Appendix for Chapter 5

F.1 Useful Lemmas

Lemma F.1. *Let Z be a matrix with rows $z_i \in \mathbb{R}^n$, for $i \in [n]$, such that $n = o(n/\log^4 n)$. Assume that $\{z_i\}_{i\in[n]}$ are independent sub-exponential random vectors with $\psi = \max_i \|z_i\|_{\psi_1}$. Let $\eta_{\mathsf{min}} = \min_i \|z_i\|_2$ and $\eta_{\mathsf{max}} = \max_i \|z_i\|_2$. Set $\xi = \psi K + K'$, and $\Delta := C\xi^2 n^{1/4} n^{3/4}$. Then, we have that*

$$\lambda_{\min}\left(ZZ^\top\right) \geq \eta_{\mathsf{min}}^2 - \Delta, \qquad \lambda_{\max}\left(ZZ^\top\right) \leq \eta_{\mathsf{max}}^2 + \Delta \tag{F.1}$$

holds with probability at least

$$1 - \exp\left(-cK\sqrt{n}\log\left(\frac{2n}{n}\right)\right) - \mathbb{P}\left(\eta_{\mathit{max}} \geq K'\sqrt{n}\right), \tag{F.2}$$

where c and C are numerical constants.

Proof. Following the notation in Adamczak et al. [2011], we define

$$B := \sup_{u\in\mathbb{R}^n:\|u\|_2=1} \left| \left\| \sum_{i=1}^n u_i z_i \right\|_2^2 - \sum_{i=1}^n u_i^2 \|z_i\|_2^2 \right|^{\frac{1}{2}}. \tag{F.3}$$

Then, for any $u \in \mathbb{R}^n$ with unit norm, we have that

$$\|Zu\|_2^2 = \left\| \sum_{i=1}^n u_i z_i \right\|_2^2 - \sum_{i=1}^n u_i^2 \|z_i\|_2^2 + \sum_{i=1}^n u_i^2 \|z_i\|_2^2 \geq \min_i \|z_i\|_2^2 - B^2, \tag{F.4}$$

which implies that

$$\lambda_{\min}\left(ZZ^\top\right) = \inf_{u\in\mathbb{R}^n:\|u\|_2=1} \|Zu\|_2^2 \geq \min_i \|z_i\|_2^2 - B^2. \tag{F.5}$$

In the same way, it can also be proven that

$$\lambda_{\max}\left(ZZ^\top\right) \leq \max_i \|z_i\|_2^2 + B^2. \tag{F.6}$$

In our case, $z_i \in \mathbb{R}^n$. In the statement of Theorem 3.2 of Adamczak et al. [2011], let's fix $r = 1$, $m = n$, and $\theta = (n/n)^{1/4} < 1/4$. Then, we have that the condition required to apply Theorem 3.2 is satisfied, *i.e.* ,

$$n \log^2 \left(2\sqrt[4]{\frac{n}{n}} \right) \leq \sqrt{nn}. \tag{F.7}$$

By combining (F.5) and (F.6) with the upper bound on B given by Theorem 3.2 of Adamczak et al. [2011], the desired result readily follows. □

Lemma F.2. *Let $\{x_i\}_{i=1}^n$ be random variables in $\mathbb{R}$ such that they all respect*

$$\mathbb{P}(|x_i| > t) < 2e^{-ct^\gamma}. \tag{F.8}$$

Then, we have

$$\max_i |x_i| \leq \log^{2/\gamma} n, \tag{F.9}$$

with probability at least $1 - 2\exp(-c \log^2 n)$, where c is a numerical constant.

Proof. For any $i \in [n]$,

$$\mathbb{P}(|x_i| > \log^{2/\gamma} n) < 2\exp(-c \log^2 n), \tag{F.10}$$

which gives

$$\mathbb{P}(\max_i |x_i| > \log^{2/\gamma} n) < n\mathbb{P}(|x_1| > \log^{2/\gamma} n) < 2\exp(\log n - c \log^2 n) < 2\exp(-c_1 \log^2 n), \tag{F.11}$$

where the second step is obtained through a union bound. This gives the desired result. □

Lemma F.3. *Let $z \sim P_Z$ be a mean-0 and Lipschitz concentrated random vector (see* (B.2)*). Consider*

$$\Gamma(z) = z \otimes z - \mathbb{E}_z \left[z \otimes z \right]. \tag{F.12}$$

Then,

$$\left\| \Gamma(z) \right\|_{\psi_1} < C. \tag{F.13}$$

where C is a numerical constant.

Proof. We have that

$$\left\| \Gamma(z) \right\|_{\psi_1} = \sup_{\|u\|_2 = 1} \left\| u^\top \Gamma(z) \right\|_{\psi_1} = \sup_{\|U\|_F = 1} \left\| z^\top U z - \mathbb{E}_x \left[z^\top U z \right] \right\|_{\psi_1}. \tag{F.14}$$

Since z is mean-0 and Lipschitz concentrated, we can apply the version of the Hanson-Wright inequality given by Theorem 2.3 in Adamczak [2015]:

$$\mathbb{P}\left(|z^\top U z - \mathbb{E}_x \left[z^\top U z \right] | > t \right) < 2\exp\left(-\frac{1}{C_1} \min\left(\frac{t^2}{\|U\|_F^2}, \frac{t}{\|U\|_{\mathsf{op}}} \right) \right), \tag{F.15}$$

where C_1 is a numerical constant. Thus, by Lemma 5.5 of Soltanolkotabi et al. [2018], and using $\|U\|_F \geq \|U\|_{\mathsf{op}}$, we conclude that

$$\left\| \Gamma(z) \right\|_{\psi_1} < C \left\| U \right\|_F = C, \tag{F.16}$$

for some numerical constant C, which gives the desired result. □

Lemma F.4. *Let $z \in \mathbb{R}^d$ be a sub-Gaussian vector such that $\|z\|_{\psi_2} = K$. Set $\Sigma = \mathbb{E}\left[zz^\top\right]$. Then,*

$$\|\Sigma\|_{\mathsf{op}} \leq CK^2, \tag{F.17}$$

where C is a numerical constant.

Proof. Let w be the unitary eigenvector associated to the maximum eigenvalue of Σ. Then,

$$\|\Sigma\|_{\mathsf{op}} = w^\top \Sigma w = \mathbb{E}\left[w^\top zz^\top w\right] = \mathbb{E}\left[(w^\top z)^2\right]. \tag{F.18}$$

Furthermore, we have that

$$\|z\|_{\psi_2} := \sup_{w' \mathsf{s.t.} \|w'\|_2 = 1} \left\|(w')^\top z\right\|_{\psi_2} \geq \left\|w^\top z\right\|_{\psi_2} \geq \frac{1}{C}\sqrt{\mathbb{E}\left[(w^\top z)^2\right]}, \tag{F.19}$$

where C is a numerical constant, and the last inequality comes from Eq. (2.15) of Vershynin [2018]. By combining (F.18) and (F.19) we get our desired result. □

Lemma F.5. *Let Z be an $n \times n$ matrix whose rows z_i are i.i.d. mean-0 sub-Gaussian random vectors in $\mathbb{R}^n$. Let $K = \|z_i\|_{\psi_2}$ the sub-Gaussian norm of each row. Then, we have*

$$\left\|ZZ^\top\right\|_{\mathsf{op}} = K^2 O\left(n + n\right), \tag{F.20}$$

with probability at least $1 - 2\exp(-cn)$, for some numerical constant c.

Proof. The result is equivalent to Lemma B.7 of Bombari et al. [2022b]. □

Lemma F.6. *Let $\{Y_i\}_{i=1}^n$ be independent real random variables such that $\|Y_i\|_{\psi_\alpha} \leq K$ for every i (see* (B.1)*), with $\alpha \in (0, 2]$. Then, for every $t > 0$ we have,*

$$\mathbb{P}\left(\sum_{i=1}^n (Y_i - \mathbb{E}[Y_i]) > t\right) < 2\exp\left(-c\left(\frac{t}{K\sqrt{n}\log^{1/\alpha} n}\right)^\alpha\right), \tag{F.21}$$

where c is an absolute constant.

Proof. Let Y be the vector containing the Y_i's in its entries. Then, an application of Cauchy–Schwarz inequality gives that the function

$$f(Y) := \frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i \tag{F.22}$$

is 1-Lipschitz.

By Lemma 5.6 of Sambale [2020], we also have that

$$\|\max_i |Y_i|\|_{\psi_\alpha} \leq CK \log^{1/\alpha} n, \tag{F.23}$$

where C is a numerical constant (that might depend on α). Thus, as f is also convex, we can apply Proposition 3.1 of Sambale [2020], obtaining

$$\mathbb{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n (Y_i - \mathbb{E}[Y_i]) > t\right) < 2\exp\left(-c\frac{t^\alpha}{K^\alpha \log n}\right). \tag{F.24}$$

Rescaling t gives the thesis. □

F.2 Proofs for Random Features

In this section, we will indicate with $X \in \mathbb{R}^{n\times d}$ the data matrix, such that its rows are sampled independently from P_X (see Assumption 5.1). We will indicate with $V \in \mathbb{R}^{k\times d}$ the random features matrix, such that $V_{ij} \sim_{\text{i.i.d.}} \mathcal{N}(0, 1/d)$. The feature vectors will be therefore indicated by $\phi(XV^\top)$, where ϕ is the activation function, applied component-wise to the pre-activations $XV^\top$. We will also use the shorthands $\Phi := \phi(XV^\top) \in \mathbb{R}^{n\times k}$ and $K := \Phi\Phi^\top \in \mathbb{R}^{n\times n}$. We also define $\tilde{\Phi} := \Phi - \mathbb{E}_X[\Phi]$ and $\tilde{K} := \tilde{\Phi}\tilde{\Phi}^\top$. Notice that, for compactness, we dropped the subscripts "RF" from these quantities, as this section will only treat the proofs related to Section 5.4. Again for the sake of compactness, we will not re-introduce such quantities in the statements or the proofs of the following lemmas.

F.2.1 Proof of Theorem 5.6

Proof of Theorem 5.6. We have

$$\|AY\|_2^2 = \|A\left(g(X) + \epsilon\right)\|_2^2 = \epsilon^\top A^\top A\epsilon + g(X)^\top A^\top A g(X) + 2g(X)^\top A^\top A\epsilon. \tag{F.25}$$

As ϵ is sub-Gaussian, we have $\|A\epsilon\|_{\psi_2} \le C\|A\|_{\text{op}}$, where C is a numerical constant. Indicating with $M := \|Ag(X)\|_2$, we can write

$$\mathbb{P}(|g(X)^\top A^\top A\epsilon| > t) < 2\exp\left(-c\frac{t^2}{M^2\|A\|_{\text{op}}^2}\right). \tag{F.26}$$

Setting $t = y\|A\|_{\text{op}} M$ in the previous equation, we readily get

$$\mathbb{P}(|g(X)^\top A^\top A\epsilon| < y\|A\|_{\text{op}} M) > 1 - 2\exp\left(-cy^2\right). \tag{F.27}$$

We will condition on this event for the rest of the proof.

On the other hand, a direct application of Theorem 6.3.2 in Vershynin [2018], gives

$$\|\|A\epsilon\|_2 - \varepsilon\|A\|_F\|_{\psi_2} \le C_1\|A\|_{\text{op}}, \tag{F.28}$$

where C_1 is an absolute constant (that depends on ε). This means that

$$\mathbb{P}\left(|\|A\epsilon\|_2 - \varepsilon\|A\|_F| > z\|A\|_{\text{op}}\right) < 2\exp\left(-c_1 z^2\right). \tag{F.29}$$

This also implies

$$\mathbb{P}\left(\|A\epsilon\|_2 \ge \varepsilon\|A\|_F/\sqrt{2}\right) > 1 - 2\exp\left(-c_2\frac{\|A\|_F^2}{\|A\|_{\text{op}}^2}\right). \tag{F.30}$$

We will condition also on this other event for the rest of the proof.

Using the triangle inequality, (F.25) gives

$$\|AY\|_2^2 \ge \epsilon^\top A^\top A\epsilon + M^2 - 2y\|A\|_{\text{op}} M \ge \epsilon^\top A^\top A\epsilon - y^2\|A\|_{\text{op}}^2 \ge \frac{\varepsilon^2}{2}\|A\|_F^2 - y^2\|A\|_{\text{op}}^2, \tag{F.31}$$

where the second inequality is obtained minimizing over M. Setting $y = \|A\|_F/(2\|A\|_{\text{op}})$ we get

$$\|AY\|_2 \ge \frac{\varepsilon}{2}\|A\|_F, \tag{F.32}$$

with probability at least $1 - \exp\left(-c_3\|A\|_F^2/\|A\|_{\text{op}}^2\right)$, which concludes the proof. □

F.2.2 Bounds on the Extremal Eigenvalues of the Kernel

Lemma F.7. *Let $v \in \mathbb{R}^d$ be distributed as $V_{j:}$, for some $j \in [k]$, independent from X. Then, we have*

$$\|\phi(Xv)\|_{\psi_2} = O\left(\sqrt{n}\right), \tag{F.33}$$

where the sub-Gaussian norm is intended over the probability space of v. This result holds with probability 1 over X.

Proof. By triangular inequality, we have

$$\|\phi(Xv)\|_{\psi_2} \leq \|\phi(Xv) - \mathbb{E}_v\left[\phi(Xv)\right]\|_{\psi_2} + \|\mathbb{E}_v\left[\phi(Xv)\right]\|_{\psi_2}. \tag{F.34}$$

Let's bound the two terms separately. As ϕ is an L-Lipschitz function, and $\sqrt{d}v$ is a Lipschitz concentrated random vector, we have

$$\|\phi(Xv) - \mathbb{E}_v\left[\phi(Xv)\right]\|_{\psi_2} \leq CL\frac{\|X\|_{\mathsf{op}}}{\sqrt{d}} \leq CL\frac{\|X\|_F}{\sqrt{d}} = CL\sqrt{n}, \tag{F.35}$$

where in the last equality we used the fact that $\|x_i\|_2 = \sqrt{d}$.

For the second term of (F.34), we have

$$\|\mathbb{E}_v\left[\phi(Xv)\right]\|_{\psi_2} \leq C\left\|\mathbb{E}_v\left[\phi(Xv)\right]\right\|_2. \tag{F.36}$$

Let $(\phi(Xv))_i$ denote the i-th component of the vector $\phi(Xv)$. As v is a Gaussian vector with covariance I/d and since $\|x_i\| = \sqrt{d}$ for all $i \in [n]$, we have $\mathbb{E}_v\left[(\phi(Xv))_i\right] = \mathbb{E}_\rho[\phi(\rho)] = O(1)$, where ρ is a standard Gaussian random variable, and the last equality is true since ϕ is Lipschitz. Thus, we have

$$\|\mathbb{E}_v\left[\phi(Xv)\right]\|_{\psi_2} = O\left(\sqrt{n}\right). \tag{F.37}$$

Putting together (F.34), (F.35), and (F.37), we get the thesis. □

Lemma F.8. *Let $v \in \mathbb{R}^d$ be distributed as $V_{j:}$, for some $j \in [k]$, independent from X. Then, we have*

$$\|\phi(Xv) - \mathbb{E}_X\left[\phi(Xv)\right]\|_{\psi_2} = O\left(\sqrt{n}\right), \tag{F.38}$$

where the sub-Gaussian norm is intended over the probability space of v. This result holds with probability 1 over X.

Proof. By triangle inequality, we have

$$\|\phi(Xv) - \mathbb{E}_X\left[\phi(Xv)\right]\|_{\psi_2} \leq \|\phi(Xv)\|_{\psi_2} + \|\mathbb{E}_X\left[\phi(Xv)\right]\|_{\psi_2}. \tag{F.39}$$

The first term is $O(\sqrt{n})$, due to Lemma F.7.

For the second term, by Jensen inequality, we have

$$\|\mathbb{E}_X\left[\phi(Xv)\right]\|_{\psi_2} \leq \mathbb{E}_X\left[\|\phi(Xv)\|_{\psi_2}\right] \leq C\mathbb{E}_X\left[\sqrt{n}\right] = O\left(\sqrt{n}\right). \tag{F.40}$$

Thus, the thesis readily follows. □

Lemma F.9. *Let $Y := (X^{*2} + (\alpha - 1)\mathbb{E}_X[X^{*2}])$ for some $\alpha \in \mathbb{R}$, where $*$ refers to the Khatri-Rao product, defined in Section 5.3. Then, we have*

$$\lambda_{\min}\left(YY^\top\right) = \Omega(d^2), \tag{F.41}$$

with probability at least $1 - \exp(-c\log^2 n)$ over X.

Proof. First, we want to compare $YY^\top$ with $\tilde{Y}\tilde{Y}^\top$, where

$$\tilde{Y} := X^{*2} - \mathbb{E}_X\left[X^{*2}\right]. \tag{F.42}$$

Let's also define $\bar{Y} := \mathbb{E}_X[X^{*2}] := \mathbf{1}\mu^\top$, where $\mathbf{1} \in \mathbb{R}^n$ denotes a vector full of ones, and $\mu := \mathbb{E}[x_1 \otimes x_1]$. We will refer to the i-th row of X with x_i. It is straightforward to verify that $Y = \tilde{Y} + \alpha\bar{Y}$. Thus, we can write

$$\begin{aligned}
YY^\top &= \tilde{Y}\tilde{Y}^\top + \alpha^2\bar{Y}\bar{Y}^\top + \alpha\tilde{Y}\bar{Y}^\top + \alpha\bar{Y}\tilde{Y}^\top \\
&= \tilde{Y}\tilde{Y}^\top + \alpha^2\left\|\mu\right\|_2^2\mathbf{1}\mathbf{1}^\top + \alpha\Lambda\mathbf{1}\mathbf{1}^\top + \alpha\mathbf{1}\mathbf{1}^\top\Lambda \\
&= \tilde{Y}\tilde{Y}^\top + \left(\alpha\left\|\mu\right\|_2\mathbf{1} + \frac{\Lambda\mathbf{1}}{\left\|\mu\right\|_2}\right)\left(\alpha\left\|\mu\right\|_2\mathbf{1} + \frac{\Lambda\mathbf{1}}{\left\|\mu\right\|_2}\right)^\top - \frac{\Lambda\mathbf{1}\mathbf{1}^\top\Lambda}{\left\|\mu\right\|_2^2} \\
&\succeq \tilde{Y}\tilde{Y}^\top - \frac{\Lambda\mathbf{1}\mathbf{1}^\top\Lambda}{\left\|\mu\right\|_2^2},
\end{aligned} \tag{F.43}$$

where we define Λ as a diagonal matrix such that $\Lambda_{ii} = \mu^\top\tilde{Y}_{i:}$.

Since $\tilde{Y}_{i:} = x_i \otimes x_i - \mathbb{E}[x_i \otimes x_i] = x_i \otimes x_i - \mu$, where x_i is mean-0 and Lipschitz concentrated, by Lemma F.3, we have that $\left\|\tilde{Y}_{i:}\right\|_{\psi_1} \leq C$, for some numerical constant C. Furthermore, we have that

$$\left|\left\|\tilde{Y}_{i:}\right\|_2 - d\right| = \left|\left\|\tilde{Y}_{i:}\right\|_2 - \left\|x_i \otimes x_i\right\|_2\right| \leq \left\|\tilde{Y}_{i:} - x_i \otimes x_i\right\|_2 = \left\|\mu\right\|_2. \tag{F.44}$$

where the second step follows from triangle inequality. We can upper-bound the RHS with the following argument:

$$\left\|\mu\right\|_2 = \left\|\mathbb{E}_{x_1}\left[x_1x_1^\top\right]\right\|_F \leq \sqrt{d}\left\|\mathbb{E}_{x_1}\left[x_1x_1^\top\right]\right\|_{\mathsf{op}} = O\left(\sqrt{d}\right), \tag{F.45}$$

where the last equality is justified by Lemma F.4, since x_i is sub-Gaussian. Then, plugging this last relation in (F.44), we have that, for all $i \in [n]$,

$$\left\|\tilde{Y}_{i:}\right\|_2 = \Theta(d) < C_y d, \tag{F.46}$$

for some numerical constant C_y.

Note that $\tilde{Y}$ has n independent rows in $\mathbb{R}^{d^2}$ such that $\left\|\tilde{Y}_{i:}\right\|_{\psi_1} \leq C$ for all i and $\min_i\left\|\tilde{Y}_{i:}\right\|_2 = \Omega(d)$. Assumption 5.4 directly implies $n = o(d^2/\log^4 d^2)$. Thus, we can apply Lemma F.1, setting $K = 1$ and $K' = C_y$ in the statement, and we get

$$\lambda_{\min}\left(\tilde{Y}\tilde{Y}^\top\right) = \Omega(d^2) - cn^{1/4}d^{3/2} = \Omega(d^2), \tag{F.47}$$

with probability at least $1 - \exp(-c\sqrt{n})$ over X, where c is a numerical constant.

To conclude the proof, we have to control the last term of (F.43). As $\Lambda_{ii} = \mu^\top \tilde{Y}_{i:}$ and $\left\|\tilde{Y}_{i:}\right\|_{\psi_1} \leq C$, we can write

$$\mathbb{P}(|\Lambda_{ii}|/\left\|\mu\right\|_2 > t) < 2\exp(-c_1 t), \tag{F.48}$$

where c_1 is a numerical constant, and the probability is intended over X. After applying Lemma F.2 with $\gamma = 1$, the last relation implies that

$$\left\|\Lambda/\left\|\mu\right\|_2\right\|_{\mathsf{op}} = O\left(\log^2 n\right), \tag{F.49}$$

with probability at least $1 - 2\exp(-c_2 \log^2 n)$ over X, where c_2 is a numerical constant.

Plugging (F.47) and (F.49) in (F.43), we get that, with probability at least $1-\exp(-c_3 \log^2 n)$,

$$\lambda_{\min}\left(YY^\top\right) \geq \lambda_{\min}\left(\tilde{Y}\tilde{Y}^\top\right) - \left\|\frac{\Lambda \mathbf{1}\mathbf{1}^\top \Lambda}{\left\|\mu\right\|_2^2}\right\|_{\mathsf{op}} = \Omega(d^2) - nO\left(\log^4 n\right) = \Omega(d^2). \tag{F.50}$$

where the last equality is again consequence of Assumption 5.4. □

Lemma F.10. *We have that*

$$\lambda_{\min}\left(\mathbb{E}_V\left[K\right]\right) = \Omega(k), \tag{F.51}$$

with probability at least $1 - \exp(-c\log^2 n)$ *over* X*, where* c *is an absolute constant.*

Proof. We have

$$\mathbb{E}_V\left[K\right] = \mathbb{E}_V\left[\sum_{i=1}^{k} \phi(XV_{i:}^\top)\phi(XV_{i:}^\top)^\top\right] = k\mathbb{E}_v\left[\phi(Xv)\phi(Xv)^\top\right] := kM, \tag{F.52}$$

where we used the shorthand v to indicate a random variable distributed as $V_{1:}$. We will also indicate with x_i the i-th row of X.

Exploiting the Hermite expansion of ϕ, we can write

$$M_{ij} = \mathbb{E}_v\left[\phi(x_i^\top v)\phi(x_j^\top v)\right] = \sum_{n=0}^{+\infty} \mu_n^2 \frac{\left(x_i^\top x_j\right)^n}{d^n} = \sum_{n=0}^{+\infty} \mu_n^2 \frac{\left[\left(X^{*n}\right)\left(X^{*n}\right)^\top\right]_{ij}}{d^n}, \tag{F.53}$$

where μ_n is the n-th Hermite coefficient of ϕ. Notice that the previous expansion was possible since $\|x_i\| = \sqrt{d}$ for all $i \in [n]$. As ϕ is non-linear, there exists $m \geq 2$ such that $\mu_m^2 > 0$. In particular, we have $M \succeq \mu_m^2 M_m$ in a p.s.d. sense, where we define

$$M_m := \frac{1}{d^m}\left(X^{*m}\right)\left(X^{*m}\right). \tag{F.54}$$

Let's consider the case $m = 2$. Then, since all the rows of X are mean-0, independent and Lipschitz concentrated, such that $\|x_i\| = \sqrt{d}$, we can apply Lemma F.9 with $\alpha = 1$ to readily get

$$\lambda_{\min}\left(M_m\right) = \Omega(1), \tag{F.55}$$

with probability at least $1 - \exp(-c\log^2 n)$ over X.

Let's now consider the case $m \geq 3$. For $i \neq j$ we can exploit again the fact that x_i is sub-Gaussian and independent from x_j, obtaining

$$\mathbb{P}\left(\left|x_i^\top x_j\right| > t\sqrt{d}\right) < \exp\left(-c_1 t^2\right). \tag{F.56}$$

Performing a union bound we also get

$$\mathbb{P}\left(\max_{i,j}\left|x_i^\top x_j\right| > t\sqrt{d}\right) < n^2 \exp\left(-c_1 t^2\right). \tag{F.57}$$

Then, by Gershgorin circle theorem, we have that

$$\lambda_{\min}\left(M_m\right) \geq 1 - \max_i \sum_{j\neq i} \frac{\left|\left(x_i^\top x_j\right)^m\right|}{d^m} \geq 1 - \frac{n}{d^m}\max_{i,j}\left|(x_i^\top x_j)^m\right| \geq 1 - \frac{nt^m}{d^{m/2}}, \tag{F.58}$$

where the last inequality holds with probability $1 - n^2\exp\left(-c_1 t^2\right)$. Setting $t = \log n$, we get

$$\lambda_{\min}\left(M_m\right) \geq 1 - \frac{n\log^m n}{d^{m/2}} = 1 - \frac{n\log^3 n}{d^{3/2}}\frac{\log^{m-3} n}{d^{m-3/2}} = 1 - o\left(\frac{n\log^3 n}{d^{3/2}}\right) = 1 - o(1), \tag{F.59}$$

where the last inequality is a consequence of Assumption 5.4. This result holds with probability at least $1 - n^2\exp\left(-c_1\log^2 n\right) \geq 1 - \exp\left(-c_2\log^2 n\right)$, and the thesis readily follows. □

Lemma F.11. *We have that*

$$\lambda_{\min}\left(K\right) = \Omega(k). \tag{F.60}$$

with probability at least $1 - \exp\left(-c\log^2 n\right)$ *over* V *and* X*, where* c *is an absolute constant.*

Proof. First we define a truncated version of Φ, which we call $\bar{\Phi}$. Over such truncated matrix, we will be able to use known concentrations results, to prove that $\lambda_{\min}\left(\bar{\Phi}\bar{\Phi}^\top\right) = \Omega(\lambda_{\min}\left(\bar{G}\right))$, where we define $\bar{G} := \mathbb{E}_V\left[\bar{\Phi}\bar{\Phi}^\top\right]$. Finally, we will prove closeness between $\bar{G}$ and G, which is analogously defined as $G := \mathbb{E}_V\left[\Phi\Phi^\top\right]$. Let's introduce the shorthand $v_i := V_{i:}$. To be precise, we define

$$\bar{\Phi}_{:j} = \phi(Xv_j)\chi\left(\|\phi(Xv_j)\|_2^2 \leq R\right), \tag{F.61}$$

where χ is the indicator function. In this case, $\chi = 1$ if $\|\phi(Xv_j)\|_2^2 \leq R$, and $\chi = 0$ otherwise. As this is a column-wise truncation, it's easy to verify that

$$\Phi\Phi^\top \succeq \bar{\Phi}\bar{\Phi}^\top. \tag{F.62}$$

Denoting with v a generic weight v_i, we can also write

$$G = k\mathbb{E}_v\left[\phi\left(Xv\right)\phi\left(Xv\right)^\top\right], \tag{F.63}$$

$$\bar{G} = k\mathbb{E}_v\left[\phi\left(Xv\right)\phi\left(Xv\right)^\top\chi\left(\|\phi\left(Xv\right)\|_2^2 \leq R\right)\right]. \tag{F.64}$$

Since all the columns of $\bar{\Phi}$ have limited norm with probability one, we can use Matrix Chernoff Inequality (see Theorem 1.1 of Tropp [2012]). In particular, for $\delta \in [0,1]$, we have,

$$\mathbb{P}\left(\lambda_{\min}\left(\bar{\Phi}\bar{\Phi}^\top\right) \leq (1-\delta)\lambda_{\min}\left(\bar{G}\right)\right) \leq n\left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\lambda_{\min}\left(\bar{G}\right)/R}, \tag{F.65}$$

where we used the fact that $\lambda_{\max}\left(\bar{\Phi}_{:j}\bar{\Phi}_{:j}^\top\right) \leq R$ and $\bar{\Phi}\bar{\Phi}^\top = \sum_j \bar{\Phi}_{:j}\bar{\Phi}_{:j}^\top$. Setting $\delta = 1/2$ in the previous equation gives

$$\mathbb{P}\left(\lambda_{\min}\left(\bar{\Phi}\bar{\Phi}^\top\right) \leq \lambda_{\min}\left(\bar{G}\right)/2\right) \leq \exp\left(-c\lambda_{\min}\left(\bar{G}\right)/R + c_1\log n\right), \tag{F.66}$$

where this probability is over V.

Finally, we prove that the matrices G and $\bar{G}$ are "close". In particular, we have

$$\begin{aligned}\left\|G-\bar{G}\right\|_{\mathsf{op}} &= k\left\|\mathbb{E}_v\left[\phi\left(Xv\right)\phi\left(Xv\right)^\top \chi\left(\left\|\phi\left(Xv\right)\right\|_2^2 > R\right)\right]\right\|_{\mathsf{op}} \\ &\leq k\mathbb{E}_v\left[\left\|\phi\left(Xv\right)\phi\left(Xv\right)^\top \chi\left(\left\|\phi\left(Xv\right)\right\|_2^2 > R\right)\right\|_{\mathsf{op}}\right] \\ &= k\mathbb{E}_v\left[\left\|\phi\left(Xv\right)\right\|_2^2 \chi\left(\left\|\phi\left(Xv\right)\right\|_2^2 > R\right)\right] \\ &= k\int_{s=0}^{\infty}\mathbb{P}_v\left(\left\|\phi\left(Xv\right)\right\|_2 \chi\left(\left\|\phi\left(Xv\right)\right\|_2^2 > R\right) > \sqrt{s}\right)\mathrm{d}s \\ &= k\int_{s=0}^{\infty}\mathbb{P}_v\left(\left\|\phi\left(Xv\right)\right\|_2 > \max(\sqrt{R},\sqrt{s})\right)\mathrm{d}s,\end{aligned} \tag{F.67}$$

where we applied Jensen inequality in the second line. Let's define $\psi := \left\|\phi\left(Xv\right)\right\|_{\psi_2}$, where the sub-Gaussian norm is intended over the probability space of v. We have

$$\begin{aligned}\left\|G-\bar{G}\right\|_{\mathsf{op}} &\leq k\int_{s=0}^{\infty}\exp\left(-c_2\frac{\max(R,s)}{\psi^2}\right)\mathrm{d}s \\ &\leq k\int_{s=0}^{\infty}\exp\left(-c_2\frac{R+s}{2\psi^2}\right)\mathrm{d}s \\ &= Ck\psi^2\exp\left(-c_2\frac{R}{2\psi^2}\right),\end{aligned} \tag{F.68}$$

where the first inequality follows from the definition of sub-Gaussian random variable, and the second one from $\max(R,s) \geq (R+s)/2$.

By Lemma F.7, we have $\psi = O\left(\sqrt{n}\right)$. Then, setting $R = k/\log^2 n$, we get

$$\left\|G-\bar{G}\right\|_{\mathsf{op}} \leq Ck\psi^2\exp\left(-c\frac{R}{2\psi^2}\right) \leq C_1 k\exp\left(-c_2\frac{k}{n\log^2 n}+\log n\right) = o(k), \tag{F.69}$$

where the last equality is a consequence of Assumption 5.3. Notice that this happens with probability 1 over v.

Due to Lemma F.10, we have that $\lambda_{\min}\left(G\right) = \Omega(k)$, with probability at least $1-\exp\left(-c_3\log^2 n\right)$ over X. Conditioning on such event, we have

$$\lambda_{\min}\left(\bar{G}\right) \geq \lambda_{\min}\left(G\right) - \left\|G-\bar{G}\right\|_{\mathsf{op}} = \Omega(k) - o(k) = \Omega(k). \tag{F.70}$$

To conclude, looking back at (F.66), we also get $\lambda_{\min}\left(\bar{\Phi}\bar{\Phi}^\top\right) = \Omega(k)$ with probability (over V) at least

$$\begin{aligned}1-\exp\left(-c\lambda_{\min}\left(\bar{G}\right)/R + c_1\log n\right) &\geq 1-\exp\left(-c_4\frac{k\log^2 n}{k}+c_1\log n\right) \\ &\geq 1-\exp\left(-c_5\log^2 n\right).\end{aligned} \tag{F.71}$$

This concludes the proof. □

Lemma F.12. *We have that*

$$\lambda_{\min}\left(\mathbb{E}_V\left[\tilde{K}\right]\right) = \Omega(k), \tag{F.72}$$

with probability at least $1-\exp(-c\log^2 n)$ *over* X*, where* c *is an absolute constant.*

Proof. Indicating with x_i the i-th row of X, and with $\tilde{\phi}(x_i^\top v) := \phi(x_i^\top v) - \mathbb{E}_X\left[\phi(x_1^\top v)\right]$, we can write

$$\mathbb{E}_V\left[\tilde{\Phi}\tilde{\Phi}^\top\right] = \mathbb{E}_V\left[\sum_{i=1}^{k}\tilde{\phi}(XV_{i:}^\top)\tilde{\phi}(XV_{i:}^\top)^\top\right] = k\mathbb{E}_v\left[\tilde{\phi}(Xv)\tilde{\phi}(Xv)^\top\right] := kM, \tag{F.73}$$

where we use the shorthand v to indicate a random variable distributed as $V_{1:}$.

Exploiting the Hermite expansion of ϕ, we can write

$$\begin{aligned}
M_{ij} &= \mathbb{E}_v\left[\tilde{\phi}(x_i^\top v)\tilde{\phi}(x_j^\top v)\right] \\
&= \mathbb{E}_v\left[\left(\phi(x_i^\top v) - \mathbb{E}_z\left[\phi(z^\top v)\right]\right)\left(\phi(x_j^\top v) - \mathbb{E}_y\left[\phi(y^\top v)\right]\right)\right] \\
&= \mathbb{E}_{vzy}\left[\phi(x_i^\top v)\phi(x_j^\top v) + \phi(z^\top v)\phi(y^\top v) - \phi(x_i^\top v)\phi(y^\top v) - \phi(x_j^\top v)\phi(z^\top v)\right] \\
&= \mathbb{E}_{zy}\left[\sum_{n=0}^{+\infty}\mu_n^2\frac{\left(x_i^\top x_j\right)^n + \left(z^\top y\right)^n - \left(x_i^\top y\right)^n - \left(x_j^\top z\right)^n}{d^n}\right],
\end{aligned} \tag{F.74}$$

where μ_n is the n-th Hermite coefficient of ϕ, and $z \sim P_X$ and $y \sim P_X$ are two independent random variables, distributed as the data and independent from them. Notice that the previous expansion was possible since $\|x_i\| = \sqrt{d}$ for all $i \in [n]$. As ϕ is non-linear, there exists $n \geq 2$ such that $\mu_n^2 > 0$. Let m denote one of such n's (i.e, m is such that $\mu_m^2 > 0$), and define ϕ^* as the function that shares the same Hermite expansion as ϕ, but with its m-th coefficient being 0. We have

$$\begin{aligned}
M_{ij} &= \mathbb{E}_{zy}\left[\sum_{n=0}^{+\infty}\mu_n^2\frac{\left(x_i^\top x_j\right)^n + \left(z^\top y\right)^n - \left(x_i^\top y\right)^n - \left(x_j^\top z\right)^n}{d^n}\right] \\
&= \mathbb{E}_{zy}\left[\sum_{n\neq m}\mu_n^2\frac{\left(x_i^\top x_j\right)^n + \left(z^\top y\right)^n - \left(x_i^\top y\right)^n - \left(x_j^\top z\right)^n}{d^n}\right] \\
&\quad + \frac{\mu_m^2}{d^m}\left(\left(x_i^\top x_j\right)^m + \mathbb{E}_{zy}\left[\left(z^\top y\right)^m - \left(x_i^\top y\right)^m - \left(x_j^\top z\right)^m\right]\right) \\
&= \mathbb{E}_v\left[\tilde{\phi}^*(x_i^\top v)\tilde{\phi}^*(x_j^\top v)\right] + \frac{\mu_m^2}{d^m}\left(\left(x_i^\top x_j\right)^m + \mathbb{E}_{zy}\left[\left(z^\top y\right)^m - \left(x_i^\top y\right)^m - \left(x_j^\top z\right)^m\right]\right) \\
&=: \mathbb{E}_v\left[\tilde{\phi}^*(Xv)\tilde{\phi}^*(Xv)^\top\right]_{ij} + \mu_m^2\left[M_m\right]_{ij}.
\end{aligned} \tag{F.75}$$

From the last equation we deduce that $M \succeq \mu_m^2 M_m$, since $\mathbb{E}_v\left[\tilde{\phi}^*(Xv)\tilde{\phi}^*(X^\top v)\right]$ is PSD. Proving the thesis becomes therefore equivalent to prove that $\lambda_{\min}\left(M_m\right) = \Omega(1)$.

Let's consider the case $m = 2$. We can write

$$\begin{aligned}
\left[M_m\right]_{ij} &= \frac{1}{d^2}\mathbb{E}_{zy}\left[\left(x_i^\top x_j - z^\top x_j + x_i^\top y - z^\top y\right)\left(x_i^\top x_j + z^\top x_j - x_i^\top y - z^\top y\right)\right] \\
&= \frac{1}{d^2}\mathbb{E}_{zy}\left[\left((x_i - z)^\top(x_j + y)\right)\left((x_i + z)^\top(x_j - y)^\top\right)\right],
\end{aligned} \tag{F.76}$$

as the cross-terms will be linear in at least one between y and z, and $\mathbb{E}[x] = \mathbb{E}[z] = 0$.

Going back in matrix notation, we can write

$$M_m = \frac{1}{d^2}\mathbb{E}_{ZY}\left[\left((X + Z)(X - Y)^\top \circ (X - Z)(X + Y)^\top\right)\right], \tag{F.77}$$

where Y and Z are $n \times d$ matrices respectively containing n copies of y and z in their rows. We can now expand this term as

$$\begin{aligned} M_m &= \frac{1}{d^2}\mathbb{E}_{ZY}\left[((X-Z)*(X+Z))\left((X+Y)*(X-Y)\right)^\top\right] \\ &= \frac{1}{d^2}\mathbb{E}_Z\left[(X-Z)*(X+Z)\right]\mathbb{E}_Y\left[(X+Y)*(X-Y)\right]^\top. \end{aligned} \tag{F.78}$$

Since $\mathbb{E}[Y] = \mathbb{E}[Z] = 0$, the cross-terms cancel out again, giving

$$\mathbb{E}_Z\left[(X-Z)*(X+Z)\right] = X*X - \mathbb{E}_Z\left[Z*Z\right] = X*X - \mathbb{E}_X\left[X*X\right], \tag{F.79}$$

where the last equality is because all the rows of Z are distributed accordingly to P_X. The same equation holds for the term with Y in (F.78). Thus, we can conclude that

$$\lambda_{\min}\left(M_m\right) = \frac{1}{d^2}\lambda_{\min}\left(TT^\top\right), \tag{F.80}$$

where we defined $T = X*X - \mathbb{E}_X\left[X*X\right]$. As all the rows of X are mean-0, independent and Lipschitz concentrated, such that $\|x_i\| = \sqrt{d}$, we can apply Lemma F.9 with $\alpha = 0$ to readily get

$$\lambda_{\min}\left(M_m\right) = \Omega(1), \tag{F.81}$$

with probability at least $1 - \exp(-c\log^2 n)$ over X.

Let's now consider the case $m \geq 3$. Since x is a sub-Gaussian random variable, we have that $\left\|\frac{x^\top y}{\sqrt{d}}\right\|_{\psi_2} = C$ in the probability space of x. Notice that this holds for all y, as $\|y\|_2 = \sqrt{d}$. Then, its m-th momentum its bounded by $m^{m/2}$, up to a multiplicative constant. As m doesn't depend by the scalings of the problem, we can simply write

$$\mathbb{E}_x\left[\left(\frac{x^\top y}{\sqrt{d}}\right)^m\right] = O\left(1\right). \tag{F.82}$$

In the same way, we can prove that

$$\mathbb{E}_x\left[\left(\frac{x^\top x_i}{\sqrt{d}}\right)^m\right] = O\left(1\right), \qquad \mathbb{E}_y\left[\left(\frac{y^\top x_j}{\sqrt{d}}\right)^m\right] = O\left(1\right). \tag{F.83}$$

Also, for $i \neq j$ we can exploit again the fact that x_i is sub-Gaussian and independent from x_j, obtaining

$$\mathbb{P}\left(\left|x_i^\top x_j\right| > t\sqrt{d}\right) < \exp\left(-ct^2\right). \tag{F.84}$$

Performing a union bound we also get

$$\mathbb{P}\left(\max_{i\neq j}\left|x_i^\top x_j\right| > t\sqrt{d}\right) < n^2\exp\left(-ct^2\right). \tag{F.85}$$

By Gershgorin circle theorem, and setting $t = \log n$, we have that

$$\begin{aligned} \lambda_{\min}\left(M_m\right) &\geq 1 - O\left(d^{-m/2}\right) - \max_i \sum_{j\neq i} \frac{\left|\left(x_i^\top x_j\right)^m + \mathbb{E}_{xy}\left[\left(x^\top y\right)^m - \left(x_i^\top y\right)^m - \left(x_j^\top x\right)^m\right]\right|}{d^m} \\ &\geq 1 - O\left(d^{-m/2}\right) - n\max_{i\neq j} \frac{\left|\left(x_i^\top x_j\right)^m + \mathbb{E}_{xy}\left[\left(x^\top y\right)^m - \left(x_i^\top y\right)^m - \left(x_j^\top x\right)^m\right]\right|}{d^m} \\ &\geq 1 - O\left(nd^{-m/2}\right) - \frac{n\log^m n}{d^{m/2}} = 1 - o(1), \end{aligned} \tag{F.86}$$

where the third inequality holds with probability at least $1 - n^2 \exp\left(-c \log^2 n\right) = 1 - \exp\left(-c_1 \log^2 n\right)$, and the last equality is a consequence of Assumption 5.4. This concludes the proof. □

Lemma F.13. *We have that*

$$\lambda_{\min}\left(\tilde{K}\right) = \Omega(k), \tag{F.87}$$

with probability at least $1 - \exp(-c \log^2 n)$ *over* V *and* X*, where* c *is an absolute constant.*

Proof. The proof follows the same exact path as Lemma F.11. The only difference is that now ψ (defined right before (F.68)) will be $\psi := \|\phi(Xv) - \mathbb{E}_X[\phi(Xv)]\|_{\psi_2}$. To successfully conclude the proof, we have to show that $\psi = O(\sqrt{n})$, which is done in Lemma F.8. □

Lemma F.14. *We have that*

$$\left\|\tilde{K}\right\|_{\mathsf{op}} = O\left(\frac{k}{d}(n+d)\right), \tag{F.88}$$

with probability at least $1 - k\exp(-cd)$ *over* V *and* X*.*

Proof. We have that

$$\tilde{K} = \tilde{\Phi}\tilde{\Phi}^\top = \sum_{i=1}^{\lceil k/d \rceil} \tilde{\Phi}^i (\tilde{\Phi}^i)^\top, \tag{F.89}$$

where we define $\tilde{\Phi}^i$ as the $n \times d$ matrix containing the columns of $\tilde{\Phi}$ between $(i-1)d+1$ and id. If $i = \lceil k/d \rceil$, we have that the number of columns is $k - d(\lceil k/d \rceil - 1) \leq d$.

Every $\tilde{\Phi}^i$ has mean-0, i.i.d. rows $\tilde{\phi}(V^i x_j)$ (in the probability space of X), where V^i contains the rows of V between $(i-1)d+1$ and id. Notice that, besides corner cases, V^i is a $d \times d$ matrix.Since the x_j are Lipschitz concentrated, we also have

$$\left\|\tilde{\phi}(V^i x_j)\right\|_{\psi_2} \leq C \left\|V^i\right\|_{\mathsf{op}} = O(1), \tag{F.90}$$

where the last equality follows from Theorem 4.4.5 of Vershynin [2018] and from the fact that both dimensions of V^i are less or equal than d, and holds with probability $1 - \exp(-cd)$ over V^i. Conditioning on such event, a straightforward application of Lemma F.5 gives

$$\left\|\tilde{\Phi}^i (\tilde{\Phi}^i)^\top\right\|_{\mathsf{op}} = O(n+d), \tag{F.91}$$

with probability at least $1 - \exp(-c_1 d)$ over X.

Thus, performing a union bound over the i-s, we get

$$\max_i \left\|\Phi^i (\tilde{\Phi}^i)^\top\right\|_{\mathsf{op}} = O(n+d), \tag{F.92}$$

with probability at least $1 - k\exp(-c_2 d)$. Conditioning over such event, we have

$$\left\|\tilde{K}\right\|_{\mathsf{op}} \leq \sum_{i=1}^{\lceil k/d \rceil} \left\|\tilde{\Phi}^i (\tilde{\Phi}^i)^\top\right\|_{\mathsf{op}} = O\left(\frac{k}{d}(n+d)\right), \tag{F.93}$$

which gives the thesis. □

F.2.3 Centering and Estimating the Norm of A(z)

In the following section it will be convenient to define $E(z) := \nabla_z \phi(Vz)^\top \phi(XV^\top)^\top$. Notice that this gives $A(z) = E(z)K^{-1}$. Furthermore, we will use the notation $\tilde{\cdot}$ to indicate quantities involving a centering with respect to the random variable X. In particular, we indicate

$$\tilde{\Phi} = \Phi - \mathbb{E}_X[\Phi], \qquad \tilde{K} = \tilde{\Phi}\tilde{\Phi}^\top, \qquad \tilde{E} := \nabla_z \phi(Vz)^\top \tilde{\Phi}^\top, \qquad \tilde{A} = \tilde{E}\tilde{K}^{-1}. \tag{F.94}$$

For the sake of compactness, we will not re-introduce any of these quantities in the statements or the proofs of the following lemmas, and we will drop the dependence on z of A, $\tilde{A}$, E and $\tilde{E}$.

Note the results of this section will be applied only when $\nu := \mathbb{E}_{x_i}[\phi(Vx_i)]$ is not the all-0 vector. Therefore, we will assume $\|\nu\|_2 > 0$ through all the proofs here. Along the section it will be convenient to recall that $\nabla_z \phi(Vz)^\top = V^\top \mathsf{diag}\left(\phi'(Vz)\right)$.

Lemma F.15. *Let $\mathbf{1} \in \mathbb{R}^n$ denote a vector full of ones. Then, the kernel matrix K can be written as*

$$K = \tilde{K} + \eta\eta^\top - \lambda\lambda^\top, \tag{F.95}$$

where

$$\eta := \|\nu\|_2 \mathbf{1} + \tilde{\Phi}\frac{\nu}{\|\nu\|_2} = \Phi\frac{\nu}{\|\nu\|_2}, \tag{F.96}$$

$$\lambda := \tilde{\Phi}\frac{\nu}{\|\nu\|_2}. \tag{F.97}$$

Proof. Note that $\Phi = \tilde{\Phi} + \mathbf{1}\nu^\top$. Then,

$$\begin{aligned} K &= \left(\tilde{\Phi} + \mathbf{1}\nu^\top\right)\left(\tilde{\Phi} + \mathbf{1}\nu^\top\right)^\top \\ &= \tilde{\Phi}\tilde{\Phi}^\top + \tilde{\Phi}\nu\mathbf{1}^\top + \mathbf{1}\nu^\top\tilde{\Phi}^\top + \|\nu\|_2^2 \mathbf{1}\mathbf{1}^\top \\ &= \tilde{\Phi}\tilde{\Phi}^\top + \left(\|\nu\|_2 \mathbf{1} + \frac{\tilde{\Phi}\nu}{\|\nu\|_2}\right)\left(\|\nu\|_2 \mathbf{1} + \frac{\tilde{\Phi}\nu}{\|\nu\|_2}\right)^\top - \frac{\tilde{\Phi}\nu\nu^\top\tilde{\Phi}^\top}{\|\nu\|_2^2} \\ &= \tilde{\Phi}\tilde{\Phi}^\top + \eta\eta^\top - \lambda\lambda^\top. \end{aligned} \tag{F.98}$$

□

Lemma F.16. *We have that*

$$\left|\left\|EK^{-1}\right\|_F - \left\|\tilde{E}K^{-1}\right\|_F\right| = O\left(\frac{\sqrt{n+d}}{d}\right), \tag{F.99}$$

with probability at least $1 - \exp(-c\log^2 n) - k\exp(-cd)$ over X and V, where c is a numerical constant.

Proof. We have

$$E = \tilde{E} + \nabla_z \phi(Vz)^\top \nu \mathbf{1}^\top. \tag{F.100}$$

By triangle inequality, we can write

$$\left|\left\|EK^{-1}\right\|_F - \left\|\tilde{E}K^{-1}\right\|_F\right| \le \left\|\nabla_z \phi(Vz)^\top \nu \mathbf{1}^\top K^{-1}\right\|_F = \left\|\nabla_z \phi(Vz)^\top \nu\right\|_2 \left\|K^{-1}\mathbf{1}\right\|_2. \tag{F.101}$$

Let's look at the two factors separately. The first one can be upper bounded as follows

$$\left\|\nabla_z \phi(Vz)^\top \nu\right\|_2 = \left\|V^\top \mathsf{diag}\left(\phi'(Vz)\right)\nu\right\|_2 \leq \|V\|_{\mathsf{op}} L \|\nu\|_2 = O\left(\sqrt{\frac{k}{d}}\|\nu\|_2\right), \tag{F.102}$$

where the first inequality holds as ϕ is L-Lipschitz, and the second equality holds with probability at least $1 - e^{-cd}$ due to (F.207). Notice that from (F.96) and (F.97), we can write

$$\mathbf{1} = \frac{\eta}{\|\nu\|_2} - \frac{\lambda}{\|\nu\|_2}. \tag{F.103}$$

By plugging this in the second factor of (F.101) and applying the triangle inequality, we have

$$\left\|K^{-1}\mathbf{1}\right\|_2 = \frac{1}{\|\nu\|_2}\left\|K^{-1}\left(\eta - \lambda\right)\right\|_2 \leq \frac{1}{\|\nu\|_2}\left(\left\|K^{-1}\eta\right\|_2 + \left\|K^{-1}\lambda\right\|_2\right). \tag{F.104}$$

Again, let's control these last two terms separately. Since we can write $\eta = \Phi\nu/\left\|\nu\right\|_2$, we get

$$\left\|K^{-1}\eta\right\|_2 = \left\|K^{-1}\Phi\nu/\left\|\nu\right\|_2\right\|_2 \leq \left\|K^{-1}\Phi\right\|_{\mathsf{op}} = \left\|\Phi^+\right\|_{\mathsf{op}} = \lambda_{\min}\left(K\right)^{-1/2} = O\left(\sqrt{\frac{1}{k}}\right), \tag{F.105}$$

where the last equality holds with probability at $1 - \exp(-c\log^2 n)$, due to Lemma F.11. We will condition on such event for the rest of the proof. For the second term, we have

$$\begin{aligned}\left\|K^{-1}\lambda\right\|_2 &= \left\|K^{-1}\tilde{\Phi}\nu/\left\|\nu\right\|_2\right\|_2 \leq \left\|K^{-1}\right\|_{\mathsf{op}}\left\|\tilde{\Phi}\right\|_{\mathsf{op}} \\ &= \lambda_{\min}\left(K\right)^{-1} O\left(\sqrt{\frac{k(n+d)}{d}}\right) = O\left(\sqrt{\frac{n+d}{kd}}\right),\end{aligned} \tag{F.106}$$

where the first equality holds due to Lemma F.14 with probability at least $1 - k\exp(-c_1 d)$. Let's condition on this event too. Therefore, putting together (F.104), (F.105) and (F.106), we have

$$\left\|K^{-1}\mathbf{1}\right\|_2 = \frac{1}{\|\nu\|_2}\left(O\left(\sqrt{\frac{1}{k}}\right) + O\left(\sqrt{\frac{n+d}{kd}}\right)\right) = \frac{1}{\|\nu\|_2} O\left(\sqrt{\frac{n+d}{kd}}\right). \tag{F.107}$$

By combining this last expression with (F.101)-(F.102), we conclude

$$\left|\left\|EK^{-1}\right\|_F - \left\|\tilde{E}K^{-1}\right\|_F\right| = O\left(\sqrt{\frac{k}{d}}\|\nu\|_2\right)\frac{1}{\|\nu\|_2} O\left(\sqrt{\frac{n+d}{kd}}\right) = O\left(\frac{\sqrt{n+d}}{d}\right), \tag{F.108}$$

which gives the thesis. □

Lemma F.17. *We have that*

$$\left|\left\|\tilde{E}K^{-1}\right\|_F - \left\|\tilde{E}\left(K + \lambda\lambda^\top\right)^{-1}\right\|_F\right| = O\left(\frac{\sqrt{n+d}}{d}\right), \tag{F.109}$$

with probability at least $1 - \exp(-c\log^2 n) - k\exp(-cd)$ *over* X *and* V*, where* c *is a numerical constant.*

Proof. A direct application of the Sherman-Morrison formula gives

$$\left(K + \lambda\lambda^\top\right)^{-1} = K^{-1} - \frac{K^{-1}\lambda\lambda^\top K^{-1}}{1 + \lambda^\top K^{-1}\lambda}. \tag{F.110}$$

By triangle inequality we know that $|\|a\|_2 - \|b\|_2| \leq \|a - b\|_2$. This implies

$$\left|\left\|\tilde{E}K^{-1}\right\|_F - \left\|\tilde{E}\left(K + \lambda\lambda^\top\right)^{-1}\right\|_F\right| \leq \left\|\tilde{E}K^{-1} - \tilde{E}\left(K + \lambda\lambda^\top\right)^{-1}\right\|_F = \left\|\tilde{E}\frac{K^{-1}\lambda\lambda^\top K^{-1}}{1 + \lambda^\top K^{-1}\lambda}\right\|_F, \tag{F.111}$$

where the last equality is a consequence of (F.110). This term can be bounded as

$$\left\|\tilde{E}\frac{K^{-1}\lambda\lambda^\top K^{-1}}{1 + \lambda^\top K^{-1}\lambda}\right\|_F \leq \left\|\tilde{E}\right\|_{\text{op}} \left\|\frac{K^{-1}\lambda\lambda^\top K^{-1}}{1 + \lambda^\top K^{-1}\lambda}\right\|_F = \left\|\tilde{E}\right\|_{\text{op}} \frac{\lambda^\top K^{-2}\lambda}{1 + \lambda^\top K^{-1}\lambda}. \tag{F.112}$$

The first factor can be bounded as follows

$$\begin{aligned}\left\|\tilde{E}\right\|_{\text{op}} &= \left\|\nabla_z\phi(Vz)^\top\tilde{\Phi}^\top\right\|_{\text{op}} \leq \left\|V^\top \text{diag}\left(\phi'(Vz)\right)\right\|_{\text{op}} \left\|\tilde{\Phi}\right\|_{\text{op}} \\ &= O\left(\sqrt{\frac{k}{d}}\right) O\left(\sqrt{\frac{k(n+d)}{d}}\right) = O\left(k\frac{\sqrt{n+d}}{d}\right),\end{aligned} \tag{F.113}$$

with probability at least $1 - k\exp(-cd)$. Here, we used the fact that ϕ is Lipschitz, and we applied Lemma F.14 and Equation (F.207). Furthermore, since $\|M\|_{\text{op}} M - M^2$ is PSD for every symmetric PSD matrix M, the following relation holds

$$\lambda^\top K^{-2}\lambda \leq \left\|K^{-1}\right\|_{\text{op}} \lambda^\top K^{-1}\lambda. \tag{F.114}$$

Thus,

$$\frac{\lambda^\top K^{-2}\lambda}{1 + \lambda^\top K^{-1}\lambda} \leq \left\|K^{-1}\right\|_{\text{op}} \frac{\lambda^\top K^{-1}\lambda}{1 + \lambda^\top K^{-1}\lambda} \leq \left\|K^{-1}\right\|_{\text{op}} = \frac{1}{\lambda_{\min}(K)} = O\left(\frac{1}{k}\right), \tag{F.115}$$

where the last equality holds with probability at least $1 - \exp(-c\log^2 n)$, because of Lemma F.11. Plugging the last equation in (F.112) together with (F.113), and comparing with (F.111), gives the desired result. □

Lemma F.18. *We have that*

$$\left|\left\|\tilde{E}\left(\tilde{K} + \eta\eta^\top\right)^{-1}\right\|_F - \left\|\tilde{E}\tilde{K}^{-1}\right\|_F\right| = O\left(\frac{\sqrt{n+d}}{d}\right), \tag{F.116}$$

with probability at least $1 - \exp(-c\log^2 n) - k\exp(-cd)$ *over* X *and* V*, where* c *is a numerical constant.*

Proof. An application of the Sherman-Morrison formula gives

$$\left(\tilde{K} + \eta\eta^\top\right)^{-1} = \tilde{K}^{-1} - \frac{\tilde{K}^{-1}\eta\eta^\top\tilde{K}^{-1}}{1 + \eta^\top\tilde{K}^{-1}\eta}. \tag{F.117}$$

This implies

$$\left|\left\|\tilde{E}\left(\tilde{K}+\eta\eta^\top\right)^{-1}\right\|_F - \left\|\tilde{E}\tilde{K}^{-1}\right\|_F\right| \leq \left\|\tilde{E}\left(\tilde{K}+\eta\eta^\top\right)^{-1} - \tilde{E}\tilde{K}^{-1}\right\|_F = \left\|\tilde{E}\frac{\tilde{K}^{-1}\eta\eta^\top\tilde{K}^{-1}}{1+\eta^\top\tilde{K}^{-1}\eta}\right\|_F, \tag{F.118}$$

where the last equality is a consequence of (F.117). This term can be bounded as

$$\left\|\tilde{E}\frac{\tilde{K}^{-1}\eta\eta^\top\tilde{K}^{-1}}{1+\eta^\top\tilde{K}^{-1}\eta}\right\|_F \leq \left\|\tilde{E}\right\|_{\mathsf{op}}\left\|\frac{\tilde{K}^{-1}\eta\eta^\top\tilde{K}^{-1}}{1+\eta^\top\tilde{K}^{-1}\eta}\right\|_F = \left\|\tilde{E}\right\|_{\mathsf{op}}\frac{\eta^\top\tilde{K}^{-2}\eta}{1+\eta^\top\tilde{K}^{-1}\eta}. \tag{F.119}$$

The first factor can be bounded as in (F.113), in Lemma F.17, giving $\left\|\tilde{E}\right\|_{\mathsf{op}} = O\left(k\frac{\sqrt{n+d}}{d}\right)$ with probability at least $1-k\exp(-cd)$. Similarly to (F.114), we can write

$$\eta^\top\tilde{K}^{-2}\eta \leq \left\|\tilde{K}^{-1}\right\|_{\mathsf{op}}\eta^\top\tilde{K}^{-1}\eta. \tag{F.120}$$

Thus,

$$\frac{\eta^\top\tilde{K}^{-2}\eta}{1+\eta^\top\tilde{K}^{-1}\eta} \leq \left\|\tilde{K}^{-1}\right\|_{\mathsf{op}}\frac{\eta^\top\tilde{K}^{-1}\eta}{1+\eta^\top\tilde{K}^{-1}\eta} \leq \left\|\tilde{K}^{-1}\right\|_{\mathsf{op}} = \frac{1}{\lambda_{\min}\left(\tilde{K}\right)} = O\left(\frac{1}{k}\right), \tag{F.121}$$

where the last equality holds with probability at least $1-\exp(-c\log^2 n)$, because of Lemma F.13. Plugging the last equation in (F.119) together with the bound on $\left\|\tilde{E}\right\|_{\mathsf{op}}$, and comparing with (F.118), gives the desired result. □

Lemma F.19. *We have that*

$$\left|\|A\|_F - \left\|\tilde{A}\right\|_F\right| = O\left(\frac{\sqrt{n+d}}{d}\right), \tag{F.122}$$

with probability at least $1-\exp(-c\log^2 n)-k\exp(-cd)$ *over* X *and* V*, where* c *is a numerical constant.*

Proof. Recalling that $A = EK^{-1}$ and $\tilde{A} = \tilde{E}\tilde{K}^{-1}$, by triangle inequality we have

$$\begin{aligned}\left|\|A\|_F - \left\|\tilde{A}\right\|_F\right| \leq{}& \left|\left\|EK^{-1}\right\|_F - \left\|\tilde{E}K^{-1}\right\|_F\right| \\ &+ \left|\left\|\tilde{E}K^{-1}\right\|_F - \left\|\tilde{E}\left(K+\lambda\lambda^\top\right)^{-1}\right\|_F\right| \\ &+ \left|\left\|\tilde{E}\left(\tilde{K}+\eta\eta^\top\right)^{-1}\right\|_F - \left\|\tilde{E}\tilde{K}^{-1}\right\|_F\right| \\ ={}& O\left(\frac{\sqrt{n+d}}{d}\right)\end{aligned} \tag{F.123}$$

where the inequality holds because of Lemma F.16, Lemma F.17, and Lemma F.18, each of them with probability at least $1-\exp(-c\log^2 n)-k\exp(-cd)$ over X and V. This gives the thesis. □

At this point, we are ready to prove Theorem 5.7.

Proof of Theorem 5.7. We have

$$\begin{aligned}\|A(z)\|_F &\leq \left\|\nabla_z\Phi(z)^\top\tilde{\Phi}^\top\tilde{K}^{-1}\right\|_F + O\left(\sqrt{n+d}/d\right)\\ &\leq \left\|\nabla_z\Phi(z)^\top\tilde{\Phi}^\top\right\|_F\left\|\tilde{K}^{-1}\right\|_{\mathsf{op}} + O\left(\sqrt{n+d}/d\right)\\ &= \lambda_{\min}^{-1}(K)\mathcal{I}(z) + O\left(\sqrt{n+d}/d\right),\end{aligned} \tag{F.124}$$

where the first inequality is justified by Lemma F.19 and holds with probability at least $1-\exp(-c\log^2 n)-k\exp(-cd)$ over X and V. For the other bound of the first equation, we have

$$\begin{aligned}\|A(z)\|_F &\geq \left\|\nabla_z\Phi(z)^\top\tilde{\Phi}^\top\tilde{K}^{-1}\right\|_F - O\left(\sqrt{n+d}/d\right)\\ &\geq \left\|\nabla_z\Phi(z)^\top\tilde{\Phi}^\top\right\|_F\lambda_{\min}\left(\tilde{K}^{-1}\right) - O\left(\sqrt{n+d}/d\right)\\ &= \lambda_{\max}^{-1}(\tilde{K})\mathcal{I}(z) - O\left(\sqrt{n+d}/d\right),\end{aligned} \tag{F.125}$$

where the first inequality is justified again by Lemma F.19, and holds with probability at least $1-\exp(-c\log^2 n)-k\exp(-cd)$ over X and V.

The second equation directly follows from Lemma F.14 and Lemma F.12. □

F.2.4 Estimating the Norm of the Interaction Matrix

The goal of this sub-Section is to estimate the value of the Frobenius norm of the Interaction Matrix $\|\mathcal{I}_{\mathsf{RF}}\|_F$, defined in (5.19). To do so, we will estimate the squared ℓ_2 norms of its columns. In particular, we define

$$\mathcal{I}_x := \left\|\nabla_z\phi(Vz)^\top\tilde{\phi}(Vx)\right\|_2. \tag{F.126}$$

Thus, we have $\|\mathcal{I}_{\mathsf{RF}}\|_F^2 = \sum_i \mathcal{I}_{x_i}^2$. In this section, with $x\sim P_X$ we refer to a generic input data, where we drop the index for compactness. Also, we will assume $z\sim P_X$, independent from the training set.

Recalling that

$$\nabla_z\phi(Vz)^\top = V^\top\mathsf{diag}\left(\phi'(Vz)\right), \tag{F.127}$$

we have

$$\mathcal{I}_x = \left\|V^\top\mathsf{diag}\left(\phi'(Vz)\right)\tilde{\phi}(Vx)\right\|_2. \tag{F.128}$$

Along the proof, the following shorthands and notations will be useful

$$\mathcal{X}_{-jl} = \sum_{m\neq j} V_{lm}x_m, \qquad \mathcal{X}_l = \sum_{m=1}^{d} V_{lm}x_m, \tag{F.129}$$

$$\mathcal{Z}_{-jl} = \sum_{m\neq j} V_{lm}z_m, \qquad \mathcal{Z}_l = \sum_{m=1}^{d} V_{lm}z_m, \tag{F.130}$$

$$I_j = \sum_{l=1}^{k} V_{lj}\phi'\left(\mathcal{Z}_l\right)\tilde{\phi}\left(\mathcal{X}_l\right), \tag{F.131}$$

which imply

$$\mathcal{I}_x^2 = \sum_{j=1}^{d} I_j^2. \tag{F.132}$$

It will also be convenient to use the following notation

$$B_{-jl} := \left(\phi''\left(\mathcal{Z}_{-jl}\right)\tilde{\phi}\left(\mathcal{X}_{-jl}\right)z_j + \phi'\left(\mathcal{Z}_{-jl}\right)\phi'\left(\mathcal{X}_{-jl}\right)x_j\right), \tag{F.133}$$

$$B_{jl} := \left(\phi''\left(\mathcal{Z}_{l}\right)\tilde{\phi}\left(\mathcal{X}_{l}\right)z_j + \phi'\left(\mathcal{Z}_{l}\right)\phi'\left(\mathcal{X}_{l}\right)x_j\right). \tag{F.134}$$

For compactness, we will not necessarily re-introduce such quantities in the statements or the proofs of the lemmas in this section.

Lemma F.20. *Indicating with c a numerical constant, the following statements hold*

$$\max_l \left|\tilde{\phi}(\mathcal{X}_l)\right| = O\left(\log k\right), \tag{F.135}$$

with probability at least $1-\exp(-c\log^2 k)$, over x and V;

$$\max_l \left|\tilde{\phi}(\mathcal{X}_{-jl})\right| = O\left(\log k\right), \tag{F.136}$$

with probability at least $1-\exp(-c\log^2 k)$, over x and $\{V_{li}\}_{i\neq j, 1\leq l\leq k}$;

$$\left|\sum_{l=1}^{k} V_{lj}\phi'\left(\mathcal{Z}_{-jl}\right)\tilde{\phi}\left(\mathcal{X}_{-jl}\right)\right| = o\left(\frac{k}{d}\right), \tag{F.137}$$

with probability at least $1-\exp(-c\log^2 k)$, over x and V.

Proof. For the first result, as ϕ is Lipschitz and centered with respect to x, its sub-Gaussian norm (in the probability space of x) is $O\left(\|V_{l:}\|_2\right) = O\left(1\right)$, where the last equality holds with probability at least $1-e^{-c_0 d}$ over $V_{l:}$, because of Theorem 3.1.1 in Vershynin [2018]. In particular, with a union bound over $V_{l:}$, we have $O\left(\max_l \|V_{l:}\|_2\right) = O\left(1\right)$ with probability at least $1-ke^{-c_0 d} \geq 1-e^{-c_1 d}$, where the last inequality is justified by Assumption 5.4. Thus, by Lemma F.2, $\max_l \left|\tilde{\phi}(\mathcal{X}_l)\right| = O\left(\log k\right)$, with probability at least $1-\exp(-c_1 d)-\exp(-c_2\log^2 k) \geq 1-\exp(-c_3\log^2 k)$, over x and V.

Notice that, with the same argument, we can prove that $\max_l \left|\tilde{\phi}(\mathcal{X}_{-jl})\right| = O\left(\log k\right)$, with probability at least $1-\exp(-c_4\log^2 k)$, over x and $\{V_{li}\}_{i\neq j, 1\leq l\leq k}$. We will condition on such event for the rest of the proof.

For the last statement, we have

$$\sum_{l=1}^{k} V_{lj}\phi'\left(\mathcal{Z}_{-jl}\right)\tilde{\phi}\left(\mathcal{X}_{-jl}\right) =: V_{:j}^{\top} u, \tag{F.138}$$

where u is a vector independent from $V_{:j}$ such that

$$\|u\|_2 \leq \sqrt{k} L \max_l \left|\tilde{\phi}\left(\mathcal{X}_{-jl}\right)\right| = O\left(\log(k)\sqrt{k}\right). \tag{F.139}$$

Since $\|V_{:j}\|_{\psi_2} \leq C/\sqrt{d}$, we have

$$\mathbb{P}\left(\left|\sum_{l=1}^{k} V_{lj}\phi'\left(\mathcal{Z}_{-jl}\right)\tilde{\phi}\left(\mathcal{X}_{-jl}\right)\right| > \frac{\sqrt{d}}{\sqrt{k}/\log^2 k}\frac{k}{d}\right) < \exp(-c_5 \log^2 k), \tag{F.140}$$

where the probability is intended over $V_{:j}$. Thus, using Assumption 5.3, we readily get the desired result. □

Lemma F.21. *We have that*

$$\left\|\mathbb{E}_x\left[\phi'\left(\mathcal{X}_{-jl}\right)x_j\right]\right\|_{\psi_2} = O\left(\frac{1}{\sqrt{d}}\right), \tag{F.141}$$

where the norm is intended over the probability space of $\{V_{li}\}_{i\neq j}$.

Proof. Let's define an auxiliary random variable $\rho \sim \mathcal{N}(0, 1/d)$, independent from any other random variable in the problem. Let's also define $V'_{l:}$ as the copy of $V_{l:}$ where we replaced the j-th component with ρ. Notice that $V'_{l:}$ is a Gaussian vector with covariance I/d. Let's also define $\mathcal{X}'_l$ in the same way, i.e., replacing V_{lj} with ρ in its expression. All the tail norms $\|\cdot\|_{\psi_r}$ in the proof will be referred to the probability space of $V'_{l:}$.

Since ϕ' is a Lipschitz function, we can write

$$\left|\mathbb{E}_x\left[\phi'\left(\mathcal{X}_{-jl}\right)x_j\right] - \mathbb{E}_x\left[\phi'\left(\mathcal{X}'_l\right)x_j\right]\right| \leq L_1\mathbb{E}_x\left[|\rho|x_j^2\right] \leq C|\rho|, \tag{F.142}$$

where we used the fact that x_j is sub-Gaussian in the last inequality, and where C is a numerical constant. Let's now look at the function

$$\varphi(v) := \mathbb{E}_x\left[\phi'\left(v^\top x\right)x_j\right]. \tag{F.143}$$

We have

$$\begin{aligned}
|\varphi(v+v') - \varphi(v)| &= \left|\mathbb{E}_x\left[\left(\phi'\left((v+v')^\top x\right) - \phi'\left(v^\top x\right)\right)x_j\right]\right| \\
&\leq \mathbb{E}_x\left[\left|\left(\phi'\left((v+v')^\top x\right) - \phi'\left(v^\top x\right)\right)\right||x_j|\right] \\
&\leq \mathbb{E}_x\left[L_1|v'^\top x||x_j|\right].
\end{aligned} \tag{F.144}$$

Since x is sub-Gaussian (and therefore also x_j is) we have that $|v'^\top x||x_j|$ is sub-exponential with norm (with respect to the probability space of x) upper bounded by $C_1\|v'\|_2$. Then, we have

$$|\varphi(v+v') - \varphi(v)| \leq \mathbb{E}_x\left[L_1|v'^\top x||x_j|\right] \leq C_2\|v'\|_2. \tag{F.145}$$

This implies that $\|\varphi\|_{\mathsf{Lip}} \leq C_2$.

As $V'_{l:}$ is Gaussian (and hence Lipschitz concentrated) with covariance I/d, we can write

$$\left\|\varphi(V'_{l:}) - \mathbb{E}_{V'_{l:}}\left[\varphi(V'_{l:})\right]\right\|_{\psi_2} = O\left(\frac{1}{\sqrt{d}}\right). \tag{F.146}$$

Furthermore,

$$\mathbb{E}_{V'_{l:}}\left[\varphi(V'_{l:})\right] = \mathbb{E}_x\left[x_j\mathbb{E}_{V'_{l:}}\left[\phi'\left(V'^\top_{l:}x\right)\right]\right] = \mathbb{E}_x\left[x_j\right]\mathbb{E}_{\rho'}\left[\phi'\left(\rho'\right)\right] = 0, \tag{F.147}$$

where ρ' is a standard Gaussian distribution independent of x. Note that the second step holds since $\|x\|_2 = \sqrt{d}$, and the last step is justified by $\mathbb{E}[x] = 0$.

This last equation implies

$$\begin{aligned}\left\|\mathbb{E}_x\left[\phi'\left(\mathcal{X}_{-jl}\right)x_j\right]\right\|_{\psi_2} &\leq \left\|\mathbb{E}_x\left[\phi'\left(\mathcal{X}_{-jl}\right)x_j\right]-\mathbb{E}_x\left[\phi'\left(\mathcal{X}'_{l}\right)x_j\right]\right\|_{\psi_2}+\left\|\mathbb{E}_x\left[\phi'\left(\mathcal{X}'_{l}\right)x_j\right]\right\|_{\psi_2}\\ &\leq C\left\|\rho\right\|_{\psi_2}+\left\|\varphi(V'_{l:})\right\|_{\psi_2}=O\left(\frac{1}{\sqrt{d}}\right),\end{aligned} \tag{F.148}$$

where the first term in the last line is justified by (F.142). In fact, the random variable on the LHS is (with probability one) upper bounded by $C|\rho|$. Then, the bound subsists also between their tail norms.

Thus,

$$\mathbb{P}_{\{V_{li}\}_{i\neq j}}\left(\left|\mathbb{E}_x\left[\phi'\left(\mathcal{X}_{-jl}\right)x_j\right]\right|>t\right)=\mathbb{P}_{V'_{l:}}\left(\left|\mathbb{E}_x\left[\phi'\left(\mathcal{X}_{-jl}\right)x_j\right]\right|>t\right)<2\exp\left(-cd\,t^2\right), \tag{F.149}$$

where with $\mathbb{P}_{\{V_{li}\}_{i\neq j}}$ we refer to the probability space of $\{V_{li}\}_{i\neq j}$. Note that the first equality is true as the statement in the probability does not depend on the value of ρ. From this, we readily get the thesis. □

Lemma F.22. *We have that*

$$\left|I_j-\sum_l V_{lj}^2\left(\phi''\left(\mathcal{Z}_{-jl}\right)\tilde{\phi}\left(\mathcal{X}_{-jl}\right)z_j+\phi'\left(\mathcal{Z}_{-jl}\right)\phi'\left(\mathcal{X}_{-jl}\right)x_j\right)\right|=o\left(\frac{k}{d}\right). \tag{F.150}$$

with probability at least $1-\exp(-c\log^2 k)$ *over* x*,* z *and* V*, where* c *is an absolute constant.*

Proof. Exploiting Taylor's Theorem, and the fact that both ϕ' and ϕ'' are Lipschitz, we can write

$$\phi'\left(\mathcal{Z}_l\right)=\phi'\left(\mathcal{Z}_{-jl}\right)+\phi''\left(\mathcal{Z}_{-jl}\right)V_{lj}z_j+a_{1lj}, \tag{F.151}$$

$$\tilde{\phi}\left(\mathcal{X}_l\right)=\phi\left(\mathcal{X}_{-jl}\right)+\phi'\left(\mathcal{X}_{-jl}\right)V_{lj}x_j+a_{2lj}-\mathbb{E}_x\left[\phi\left(\mathcal{X}_{-jl}\right)+\phi'\left(\mathcal{X}_{-jl}\right)V_{lj}x_j+a_{2lj}\right], \tag{F.152}$$

where $|a_{1lj}|=O\left(V_{lj}^2z_j^2\right)$ and $|a_{2lj}|=O\left(V_{lj}^2x_j^2\right)$. In the previous two equations, we expanded ϕ' around $\mathcal{Z}_{-jl}$ and ϕ around $\mathcal{X}_{-jl}$. For the rest of the proof (except from when expectations with respect to x are taken) we condition on the events that both $|x_j|$ and $|z_j|$ are $O\left(\log k\right)$. As they are sub-Gaussian, this happens with probability at least $1-e^{-c_1\log^2 k}$ over x and z.

Invoking Lemma F.6, and setting $t=k/d^{1/\alpha}$, we can write (for $\alpha\in(0,2]$)

$$\sum_{l=1}^{k}|V_{lj}|^{2/\alpha}=O\left(\frac{k}{d^{1/\alpha}}\right), \tag{F.153}$$

with probability at least $1-\exp(-c_2k^{\alpha/2}/\log k)$. This is true since $\mathbb{E}_\rho\left[\rho^{2/\alpha}\right]=C_\alpha$, where ρ is a standard Gaussian random variable. Thus, with probability at least $1-\exp(-c_3k^{1/5}/\log k)$ over V, we jointly have

$$\sum_{l=1}^{k}|V_{lj}|^2=O\left(\frac{k}{d}\right),\qquad\sum_{l=1}^{k}|V_{lj}|^3=O\left(\frac{k}{d^{3/2}}\right),\qquad\sum_{l=1}^{k}|V_{lj}|^5=O\left(\frac{k}{d^{5/2}}\right). \tag{F.154}$$

We will condition on these events too. It will also be convenient to condition on the high probability events described by Lemma F.20. Thus, our discussion is now restricted to a probability space over x, z and V of measure at least $1-e^{-c_4\log^2 k}-e^{-c_5d}$.

We are now ready to estimate the cross terms of the product $V_{lj}\phi'\left(\mathcal{Z}_l\right)\tilde{\phi}\left(\mathcal{X}_l\right)$, exploiting the expansions in (F.151) and (F.152). Let's control the sums over l of all of them separately.

(a) $\phi'(\mathcal{Z}_l)(a_{2lj} - \mathbb{E}_x[a_{2lj}])$

$$\begin{aligned}\left|\sum_{l=1}^{k} V_{lj}\phi'(\mathcal{Z}_l)(a_{2lj} - \mathbb{E}_x[a_{2lj}])\right| &\leq C_1 \sum_{l=1}^{k} |V_{lj}| L(|a_{2lj}| + \mathbb{E}_x[|a_{2lj}|]) \\ &\leq C_2\left(x_j^2 + 1\right)\sum_{l=1}^{k}|V_{lj}|^3 = O\left(\frac{k\log^2 k}{d^{3/2}}\right) = o\left(\frac{k}{d}\right),\end{aligned} \tag{F.155}$$

where the second inequality is justified by the fact that x_j is sub-Gaussian, and the last equality by Assumption 5.4.

(b) $\tilde{\phi}(\mathcal{X}_l)\, a_{1jl}$

$$\begin{aligned}\left|\sum_{l=1}^{k} V_{lj}\tilde{\phi}(\mathcal{X}_l)\, a_{1jl}\right| &\leq C_3 z_j^2 \sum_{l=1}^{k}|V_{lj}|^3\left|\tilde{\phi}(\mathcal{X}_l)\right| \leq C_3 z_j^2 \max_l\left|\tilde{\phi}(\mathcal{X}_l)\right|\sum_{l=1}^{k}|V_{lj}|^3 \\ &= O\left(\log^2(k)\log(k)\frac{k}{d^{3/2}}\right) = o\left(\frac{k}{d}\right),\end{aligned} \tag{F.156}$$

where the second to last equality follows from Lemma F.20, and the last from Assumption 5.4.

(c) $a_{1lj}(a_{2lj} - \mathbb{E}_x[a_{2lj}])$ (as we counted it twice)

$$\left|\sum_{l=1}^{k} V_{lj}a_{1lj}(a_{2lj} - \mathbb{E}_x[a_{2lj}])\right| \leq C_4\left(x_j^2+1\right)z_j^2\sum_{l=1}^{k}|V_{lj}|^5 = O\left(\log^4(k)\frac{k}{d^{5/2}}\right) = o\left(\frac{k}{d}\right). \tag{F.157}$$

(d) $\phi'(\mathcal{Z}_{-jl})\,\tilde{\phi}(\mathcal{X}_{-jl})$

$$\left|\sum_{l=1}^{k} V_{lj}\phi'(\mathcal{Z}_{-jl})\,\tilde{\phi}(\mathcal{X}_{-jl})\right| = o\left(\frac{k}{d}\right), \tag{F.158}$$

as it directly follows from Lemma F.20.

(e) $(\phi'(\mathcal{Z}_{-jl}) + \phi''(\mathcal{Z}_{-jl})\, V_{lj}z_j)\,\mathbb{E}_x[\phi'(\mathcal{X}_{-jl})\, V_{lj}x_j]$

$$\begin{aligned}&\left|\sum_{l=1}^{k} V_{lj}\left(\phi'(\mathcal{Z}_{-jl}) + \phi''(\mathcal{Z}_{-jl})\, V_{lj}z_j\right)V_{lj}\mathbb{E}_x[\phi'(\mathcal{X}_{-jl})\, x_j]\right| \\ &\quad\leq \max_l|\phi'(\mathcal{Z}_{-jl}) + \phi''(\mathcal{Z}_{-jl})\, V_{lj}z_j|\max_l|\mathbb{E}_x[\phi'(\mathcal{X}_{-jl})\, x_j]|\sum_{l=1}^{k}|V_{lj}|^2 \\ &\quad\leq (L + L_1\max_l|V_{lj}|\,|z_j|)\max_l|\mathbb{E}_x[\phi'(\mathcal{X}_{-jl})\, x_j]|\sum_{l=1}^{k}|V_{lj}|^2 \\ &\quad= O\left(\left(1 + \frac{\log^2 k}{\sqrt{d}}\right)\frac{\log k}{\sqrt{d}}\right)O\left(\frac{k}{d}\right) = o\left(\frac{k}{d}\right),\end{aligned} \tag{F.159}$$

where the second to last equality comes from the fact that both $\|V_{lj}\|_{\psi_2}$ and $\|\mathbb{E}_x[\phi'(\mathcal{X}_{-jl})\, x_j]\|_{\psi_2}$ are $O\left(\frac{1}{\sqrt{d}}\right)$ (see Lemma F.21) and holds with probability at least $1 - e^{-c_6\log^2 k}$ over V. The last equality is true for Assumption 5.4.

(f) $\phi''\left(\mathcal{Z}_{-jl}\right) V_{lj} z_j \phi'\left(\mathcal{X}_{-jl}\right) V_{lj} x_j$ As ϕ and ϕ' are Lipschitz

$$\left|\sum_{l=1}^{k} V_{lj}\phi''\left(\mathcal{Z}_{-jl}\right) V_{lj} z_j \phi'\left(\mathcal{X}_{-jl}\right) V_{lj} x_j\right| \leq LL_1|x_j z_j| \sum_{l=1}^{k} |V_{lj}|^3 = O\left(\log^2 k \frac{k}{d^{3/2}}\right) = o\left(\frac{k}{d}\right). \tag{F.160}$$

Then, we are left only with two cross terms, and we can write

$$I_j = \sum_l V_{lj}^2 \left(\phi''\left(\mathcal{Z}_{-jl}\right) \tilde{\phi}\left(\mathcal{X}_{-jl}\right) z_j + \phi'\left(\mathcal{Z}_{-jl}\right) \phi'\left(\mathcal{X}_{-jl}\right) x_j\right) + \text{rest}, \tag{F.161}$$

where |rest| can be upper-bounded by the sum of the terms (a)-(f), which gives

$$\left|I_j - \sum_l V_{lj}^2 \left(\phi''\left(\mathcal{Z}_{-jl}\right) \tilde{\phi}\left(\mathcal{X}_{-jl}\right) z_j + \phi'\left(\mathcal{Z}_{-jl}\right) \phi'\left(\mathcal{X}_{-jl}\right) x_j\right)\right| = o\left(\frac{k}{d}\right) \tag{F.162}$$

with probability at least $1 - \exp(-c_5 d) - \exp(-c_7 \log^2 k)$ over x, z and V. Hence, the thesis follows after using Assumption 5.4. □

Lemma F.23. *We have that*

$$\left|\sum_{l=1}^{k} V_{lj}^2 B_{-jl} - \frac{1}{d}\sum_{l=1}^{k} B_{-jl}\right| = o\left(\frac{k}{d}\right), \tag{F.163}$$

with probability at least $1 - \exp(-c \log^2 k)$ *over* x, z *and* V, *where* c *is an absolute constant.*

Proof. We have

$$\sum_{l=1}^{k} V_{lj}^2 B_{-jl} = \sum_{l=1}^{k} \left(V_{lj}^2 - \mathbb{E}_V\left[V_{lj}^2\right]\right) B_{-jl} + \frac{1}{d}\sum_{l=1}^{k} B_{-jl}, \tag{F.164}$$

since $\mathbb{E}_V\left[V_{lj}^2\right] = 1/d$ for all l, j.

We call $v_l := V_{lj}^2 - \mathbb{E}_V\left[V_{lj}^2\right]$. Then, $v = (v_1, \ldots, v_k)$ is a mean-0 vector with independent and sub-exponential entries with norm $O\left(1/d\right)$. This implies $\|v\|_{\psi_1} = O\left(1/d\right)$. Thus,

$$\mathbb{P}_v\left(|v^\top u| > t\right) < 2\exp\left(-c\frac{dt}{\|u\|_2}\right), \tag{F.165}$$

for every vector u independent from v. If we define $u_l := B_{-jl}$, we have

$$\|u\|_2 \leq \sqrt{k}\left(\max_l |B_{-jl}|\right) \leq \sqrt{k}\left(L^2|x_j| + L_1|z_j| \max_l \left|\tilde{\phi}(\mathcal{X}_{-jl})\right|\right) = O\left(\log^2(k)\sqrt{k}\right), \tag{F.166}$$

where the last equality comes from the fact that x_j and z_j are sub-Gaussian, and from the second statement in Lemma F.20. This holds with probability at least $1 - e^{-c_1 d} - e^{-c_2 \log^2 k}$ over x, z and $\{V_{li}\}_{i\neq j, 1\leq l\leq k}$, and we will condition on this event.

Then, merging this last result with (F.165) and setting $t = \log^4 k\sqrt{k}/d$ we have

$$\mathbb{P}_v\left(|v^\top u| > \frac{\log^4 k\sqrt{k}}{d}\right) < 2\exp\left(-c_3 \log^2 k\right). \tag{F.167}$$

Performing a union bound gives the thesis. □

Lemma F.24. *We have that*

$$\left|\sum_{l=1}^{k} B_{-jl} - \sum_{l=1}^{k} B_{jl}\right| = o(k), \tag{F.168}$$

with probability at least at least $1 - \exp(-c\log^2 k)$ *over* x*,* z *and* V*, where* c *is an absolute constant.*

Proof. Let's condition on

$$|x_j| = O\left(\log k\right), \qquad |z_j| = O\left(\log k\right), \qquad \max_l |V_{lj}| = O\left(\frac{\log k}{\sqrt{d}}\right). \tag{F.169}$$

These events happen with probability at least $1-\exp(-c\log^2 k)$, as $\|x_j\|_{\psi_2} = O\left(1\right)$, $\|z_j\|_{\psi_2} = O\left(1\right)$ and $\|V_{jl}\|_{\psi_2} = O\left(1/\sqrt{d}\right)$. The probability is over x, z and V. Also, we have $\max_l \left|\tilde{\phi}(\mathcal{X}_{-jl})\right| = O\left(\log k\right)$ with probability at least $1 - e^{-c_1 \log^2 k}$ over x, z and V, for the second statement in Lemma F.20. Let's condition on these high probability events for the rest of the proof, whenever we do not take expectations with respect to x.

Exploiting that ϕ, ϕ' and ϕ'' are all Lipschitz,

$$\max_l \left|\phi''\left(\mathcal{Z}_{-jl}\right) - \phi''\left(\mathcal{Z}_l\right)\right| = O\left(\frac{\log^2 k}{\sqrt{d}}\right), \tag{F.170}$$

$$\begin{aligned}&\max_l \left(\left|\phi\left(\mathcal{X}_{-jl}\right) - \phi\left(\mathcal{X}_{-jl}\right)\right| + \left|\mathbb{E}_x\left[\phi\left(\mathcal{X}_{-jl}\right) - \phi\left(\mathcal{X}_{-jl}\right)\right]\right|\right)\\ &\quad\leq O\left(\frac{\log^2 k}{\sqrt{d}}\right) + \max_l |V_{lj}|\mathbb{E}_x[|x_j|] = O\left(\frac{\log^2 k}{\sqrt{d}}\right),\end{aligned} \tag{F.171}$$

$$\max_l \left|\phi'\left(\mathcal{Z}_{-jl}\right) - \phi'\left(\mathcal{Z}_l\right)\right| = O\left(\frac{\log^2 k}{\sqrt{d}}\right), \tag{F.172}$$

$$\max_l \left|\phi'\left(\mathcal{X}_{-jl}\right) - \phi'\left(\mathcal{X}_{-jl}\right)\right| = O\left(\frac{\log^2 k}{\sqrt{d}}\right). \tag{F.173}$$

Then, controlling all the cross terms, we get

$$\max_l \left|B_{-jl} - B_{jl}\right| = O\left(\frac{\log^4 k}{\sqrt{d}}\right), \tag{F.174}$$

where this scaling is justified by the leading term

$$\max_l \left|\phi''\left(\mathcal{Z}_{-jl}\right) - \phi''\left(\mathcal{Z}_l\right)\right| \max_l \left|\tilde{\phi}(\mathcal{X}_{-jl})\right| |z_j| = O\left(\frac{\log^2 k}{\sqrt{d}} \log k \log k\right). \tag{F.175}$$

Thus, we can conclude

$$\left|\sum_{l=1}^{k} B_{-jl} - \sum_{l=1}^{k} B_{jl}\right| \leq k \max_l \left|B_{-jl} - B_{jl}\right| = O\left(\frac{k\log^4 k}{\sqrt{d}}\right) = o(k), \tag{F.176}$$

because of Assumption 5.4. This gives the thesis. □

Lemma F.25. *We have that*

$$\left|\sum_{l=1}^{k} B_{jl} - k\,\mathbb{E}_V[B_{j1}]\right| = o(k), \tag{F.177}$$

with probability at least $1-\exp(-c\log^2 k)$ *over* x, z *and* V.

Proof. Since x_j and z_j are sub-Gaussian, we know that $|x_j|, |z_j| = O(\log k)$ with probability at least $1-\exp(-c\log^2 k)$ over x and z. We will condition on such event for the rest of the proof. Note that, once we fix the values of x and z, the B_{jl}'s are i.i.d. random variables in the probability space of V. Thus,

$$\sum_{l=1}^{k} B_{jl} = k\,\mathbb{E}_V[B_{j1}] + \sum_{l=1}^{k}\left(B_{jl} - \mathbb{E}_V[B_{jl}]\right). \tag{F.178}$$

Until the end of the proof, we only refer to the probability space of V.

We have

$$\left\|\tilde{\phi}\left(\mathcal{X}_l\right)\right\|_{\psi_2} \le \left\|\phi\left(\mathcal{X}_l\right)\right\|_{\psi_2} + \left\|\mathbb{E}_x\left[\phi\left(\mathcal{X}_l\right)\right]\right\|_{\psi_2} \le C\left\|x\right\|_2 \frac{1}{\sqrt{d}} + \mathbb{E}_x\left[\left\|\phi\left(\mathcal{X}_l\right)\right\|_{\psi_2}\right] \le 2C, \tag{F.179}$$

where the second inequality is true because $\left\|x^\top V_{l:}\right\|_{\psi_2} \le C_1\left\|x\right\|\frac{1}{\sqrt{d}}$ and because ϕ is Lipschitz.

Since ϕ and ϕ' are both Lipschitz, we have that also $\left\|\phi''\left(\mathcal{Z}_l\right)\right\|_{\psi_2}$, $\left\|\phi'\left(\mathcal{Z}_l\right)\right\|_{\psi_2}$ and $\left\|\phi'\left(\mathcal{X}_l\right)\right\|_{\psi_2}$ are upper bounded by a numerical constant. Thus, we have that

$$\left\|B_{jl}\right\|_{\psi_1} \le \log k. \tag{F.180}$$

An application of Bernstein inequality (cf. Theorem 2.8.1. in Vershynin [2018]) gives that

$$\mathbb{P}\left(\left|\sum_{l=1}^{k}\left(B_{jl} - \mathbb{E}_V[B_{jl}]\right)\right| > \sqrt{k}\log^2 k\right) < \exp(-c_1\log^2 k), \tag{F.181}$$

and the thesis readily follows. □

Lemma F.26. *Indicating with* ρ *a standard Gaussian random variable, we have that*

$$\left|\mathbb{E}_V[B_{j1}] - x_j\mathbb{E}_\rho^2[\phi'(\rho)]\right| = o(1), \tag{F.182}$$

with probability at least $1-\exp(-c\log^2 k)$ *over* x *and* z.

Proof. Let's define the vector z' as follows

$$z' = \frac{\sqrt{d}\left(I - \frac{xx^\top}{d}\right)z}{\left\|\left(I - \frac{xx^\top}{d}\right)z\right\|_2}. \tag{F.183}$$

Notice that, by construction, $x^\top z' = 0$, and $\left\|z'\right\|_2 = \sqrt{d}$. Also, consider a vector y orthogonal to both z and x. Then, a fast computation returns $y^\top z' = 0$. This means that z' is the vector

on the $\sqrt{d}$-sphere, lying on the same plane of z and x, orthogonal to x. Thus, we can easily compute

$$\frac{\left|z^\top z'\right|}{d} = \sqrt{1 - \left(\frac{x^\top z}{d}\right)^2} \geq 1 - \left(\frac{x^\top z}{d}\right)^2, \tag{F.184}$$

where the last inequality derives from $\sqrt{1-a} \geq 1 - a$ for $a \in [0, 1]$. Then,

$$\|z - z'\|_2^2 = \|z\|_2^2 + \|z\|_2^2 - 2z^\top z' \leq 2d\left(1 - \left(1 - \left(\frac{x^\top z}{d}\right)^2\right)\right) = 2\frac{\left(x^\top z\right)^2}{d}. \tag{F.185}$$

As x and z are both sub-Gaussian, mean-0 vectors, with ℓ_2 norm equal to $\sqrt{d}$, we have that

$$\mathbb{P}\left(\|z - z'\|_2 > t\right) \leq \mathbb{P}\left(|x^\top z| > \sqrt{d}t/\sqrt{2}\right) < \exp(-ct^2), \tag{F.186}$$

where c is an absolute constant. Here the probability is referred to the space of x, for a fixed z. Thus, $\|z - z'\|_2$ is sub-Gaussian.

Let's now introduce the shorthands $v := V_{1j}$ and $\mathcal{Z}_1' = v^\top z'$. We have

$$\left|\mathbb{E}_v\left[\phi''(\mathcal{Z}_1)\phi(\mathcal{X}_1)\right] - \mathbb{E}_v\left[\phi''(\mathcal{Z}_1')\phi(\mathcal{X}_1)\right]\right| \leq L_2\mathbb{E}_v\left[|v^\top(z - z')||\phi(\mathcal{X}_1)|\right]. \tag{F.187}$$

Since ϕ is Lipschitz, $\phi(\mathcal{X}_1)$ is sub-Gaussian with respect to the probability space of v. Since z and z' are independent from v, we deduce that

$$\left\|v^\top(z - z')\right\|_{\psi_2} \leq \|z - z'\|_2 \frac{C}{\sqrt{d}} = O\left(\frac{\log k}{\sqrt{d}}\right), \tag{F.188}$$

with probability at least $1 - \exp(-c\log^2 k)$ over x. If we condition on such event, this implies that the expectation in (F.187) can be bounded as

$$\mathbb{E}_v\left[|v^\top(z - z')||\phi(\mathcal{X}_1)|\right] = O\left(\frac{\log k}{\sqrt{d}}\right). \tag{F.189}$$

This is true since we are computing the first moment of a sub-exponential random variable.

A similar computation gives

$$\begin{aligned}&\left|\mathbb{E}_v\left[\phi''(\mathcal{Z}_1)\mathbb{E}_x[\phi(\mathcal{X}_1)]\right] - \mathbb{E}_{vx}\left[\phi''(\mathcal{Z}_1')\phi(\mathcal{X}_1)\right]\right| \\ &\quad\leq L_2\mathbb{E}_x\mathbb{E}_v\left[|v^\top(z - z')||\phi(\mathcal{X}_1)|\right] \leq C\frac{\mathbb{E}_x\left[\|z - z'\|_2\right]}{\sqrt{d}} = O\left(\frac{1}{\sqrt{d}}\right),\end{aligned} \tag{F.190}$$

where the last step is a direct consequence of (F.186). Notice that this, together with the previous argument between (F.187) and (F.189), gives

$$\left|\mathbb{E}_v\left[\phi''(\mathcal{Z}_1)\tilde{\phi}(\mathcal{X}_1)\right] - \left(\mathbb{E}_v\left[\phi''(\mathcal{Z}_1')\phi(\mathcal{X}_1)\right] - \mathbb{E}_{vx}\left[\phi''(\mathcal{Z}_1')\phi(\mathcal{X}_1)\right]\right)\right| = O\left(\frac{\log k}{d}\right). \tag{F.191}$$

With an equivalent argument of the one between (F.187) and (F.189), we can also show that

$$\left|\mathbb{E}_v\left[\phi'(\mathcal{Z}_1)\phi'(\mathcal{X}_1)\right] - \mathbb{E}_v\left[\phi'(\mathcal{Z}_1')\phi'(\mathcal{X}_1)\right]\right| = O\left(\frac{\log k}{\sqrt{d}}\right), \tag{F.192}$$

with probability at least $1 - \exp(-c \log^2 k)$ over x.

At this point, since $x^\top z' = 0$, we have have that (in the probability space of V), $\mathcal{X}_1$ and $\mathcal{Z}_1'$ can be treated as independent standard Gaussian random variables ρ_1 and ρ_2. We therefore have

$$\mathbb{E}_v \left[\phi''(\mathcal{Z}_1')\phi(\mathcal{X}_1)\right] = \mathbb{E}_{\rho_1\rho_2} \left[\phi''(\rho_1)\phi(\rho_2)\right] = \mathbb{E}_{\rho_1} \left[\phi''(\rho_1)\right] \mathbb{E}_{\rho_2} \left[\phi(\rho_2)\right], \tag{F.193}$$

$$\mathbb{E}_{vx} \left[\phi''(\mathcal{Z}_1')\phi(\mathcal{X}_1)\right] = \mathbb{E}_x \left[\mathbb{E}_{\rho_1\rho_2} \left[\phi''(\rho_1)\phi(\rho_2)\right]\right] = \mathbb{E}_{\rho_1} \left[\phi''(\rho_1)\right] \mathbb{E}_{\rho_2} \left[\phi(\rho_2)\right], \tag{F.194}$$

$$\mathbb{E}_v \left[\phi'(\mathcal{Z}_1')\phi'(\mathcal{X}_1)\right] = \mathbb{E}_{\rho_1\rho_2} \left[\phi'(\rho_1)\phi'(\rho_2)\right] = \mathbb{E}_{\rho_1}^2 \left[\phi'(\rho_1)\right]. \tag{F.195}$$

Now, merging (F.191) with (F.193) and (F.194), we get

$$\begin{aligned} &\left|\mathbb{E}_v \left[\phi''(\mathcal{Z}_1)\tilde{\phi}(\mathcal{X}_1)\right] - \left(\mathbb{E}_{\rho_1} \left[\phi''(\rho_1)\right] \mathbb{E}_{\rho_2} \left[\phi(\rho_2)\right] - \mathbb{E}_{\rho_1} \left[\phi''(\rho_1)\right] \mathbb{E}_{\rho_2} \left[\phi(\rho_2)\right]\right)\right| \\ &= \left|\mathbb{E}_v \left[\phi''(\mathcal{Z}_1)\phi(\mathcal{X}_1)\right]\right| = O\left(\frac{\log k}{\sqrt{d}}\right). \end{aligned} \tag{F.196}$$

While merging (F.192) with (F.195) returns

$$\left|\mathbb{E}_v \left[\phi'(\mathcal{Z}_1)\phi'(\mathcal{X}_1)\right] - \mathbb{E}_{\rho_1}^2 \left[\phi'(\rho_1)\right]\right| = O\left(\frac{\log k}{\sqrt{d}}\right). \tag{F.197}$$

The previous statements hold with probability at least $1 - \exp(-c_1 \log^2 k)$ over x.

Thus, conditioning on $|x_j| = O(\log k)$ and $|z_j| = O(\log k)$, we get

$$\left|\mathbb{E}_V[B_{j1}] - x_j \mathbb{E}_\rho^2[\phi'(\rho)]\right| = O\left(\frac{\log^2 k}{\sqrt{d}}\right) = o(1), \tag{F.198}$$

with probability at least $1 - \exp(-c_2 \log^2 k)$ over x and z. This readily gives the thesis. □

At this point, we are ready to prove Theorem 5.8.

Proof of Theorem 5.8. Let's focus on $\mathcal{I}_{x_1}^2 = \sum_{j=1}^d I_j^2$, where with a slight abuse of notation we now refer to I_j as the quantity defined in (F.131), with the variable x_1 instead of x. All the results from the previous lemmas hold as $x_1 \sim P_X$. We have, by triangle inequality,

$$\begin{aligned} \left|I_j - x_j \mathbb{E}_\rho^2[\phi'(\rho)]\frac{k}{d}\right| &\leq \left|I_j - \sum_{l=1}^k V_{lj}^2 B_{-jl}\right| + \left|\sum_{l=1}^k V_{lj}^2 B_{-jl} - \frac{1}{d}\sum_{l=1}^k B_{-jl}\right| \\ &\quad + \frac{1}{d}\left|\sum_{l=1}^k B_{-jl} - \sum_{l=1}^k B_{jl}\right| + \frac{1}{d}\left|\sum_{l=1}^k B_{jl} - k\,\mathbb{E}_V[B_{j1}]\right| + \frac{k}{d}\left|\mathbb{E}_V[B_{j1}] - x_j \mathbb{E}_\rho^2[\phi'(\rho)]\right| \\ &= o\left(\frac{k}{d}\right) + o\left(\frac{k}{d}\right) + \frac{1}{d}o(k) + \frac{1}{d}o(k) + \frac{k}{d}o(1) = o\left(\frac{k}{d}\right), \end{aligned} \tag{F.199}$$

where the five terms in the third line are justified respectively by Lemmas F.22, F.23, F.24, F.25 and F.26, and jointly hold with probability at least $1 - \exp(-c_2 \log^2 k)$ over x_1 and z and V.

This holds for every $1 \leq j \leq d$ with probability at least $1 - d\exp(-c_2 \log^2 k) \geq 1 - \exp(-c_3 \log^2 k)$. Thus, we can write

$$I_{x_1} \geq \sqrt{\sum_{j=1}^{d} \left(|x_j| \mathbb{E}_\rho^2[\phi'(\rho)] \frac{k}{d} - o\left(\frac{k}{d}\right) \right)^2} \geq \sqrt{\mathbb{E}_\rho^4[\phi'(\rho)] \frac{k^2}{d^2} \sum_{j=1}^{d} x_j^2 - o\left(\frac{k^2}{d}\right) - o\left(\frac{k^2 \sum_j |x_j|}{d^2}\right)}. \tag{F.200}$$

By Cauchy-Schwartz, we have that $\sum_j |x_j| \leq \sqrt{d \sum_j x_j^2} = d$. If $\mathbb{E}_\rho[\phi'(\rho)] \neq 0$, we have

$$I_{x_1} \geq \sqrt{\mathbb{E}_\rho^4[\phi'(\rho)] \frac{k^2}{d} - o\left(\frac{k^2}{d}\right)} = \mathbb{E}_\rho^2[\phi'(\rho)] \frac{k}{\sqrt{d}} \sqrt{1 - o(1)} = \mathbb{E}_\rho^2[\phi'(\rho)] \frac{k}{\sqrt{d}} - o\left(\frac{k}{\sqrt{d}}\right). \tag{F.201}$$

Following the same strategy, we can also prove that

$$I_{x_1} \leq \mathbb{E}_\rho^2[\phi'(\rho)] \frac{k}{\sqrt{d}} + o\left(\frac{k}{\sqrt{d}}\right), \tag{F.202}$$

which finally gives

$$\left| I_{x_1} - \mathbb{E}_\rho^2[\phi'(\rho)] \frac{k}{\sqrt{d}} \right| = o\left(\frac{k}{\sqrt{d}}\right). \tag{F.203}$$

Notice that we easily retrieve the same result also in the case $\mathbb{E}_\rho[\phi'(\rho)] = 0$.

Performing now a union bound over all the data, and recalling that $\|\mathcal{I}_{\mathsf{RF}}\|_F = \sqrt{\sum_{i=1}^n I_{x_i}^2} \geq \sqrt{n} \min_i I_{x_i}$ we have

$$\|\mathcal{I}_{\mathsf{RF}}\|_F \geq \mathbb{E}_\rho^2[\phi'(\rho)] \frac{k\sqrt{n}}{\sqrt{d}} - o\left(\frac{k\sqrt{n}}{\sqrt{d}}\right), \tag{F.204}$$

with probability at least $1 - n\exp(-c_3 \log^2 k) \geq 1 - \exp(-c_4 \log^2 k)$ over X and z and V. The same argument proves also the upper bound, and the thesis readily follows. □

F.2.5 Proof of Theorem 5.5

First, we state and prove an auxiliary result upper bounding $\|A\|_{\mathsf{op}}$.

Lemma F.27. *Let $A = \nabla_z \phi(Vz)^\top \phi(XV^\top)^\top K^{-1}$. Then, we have*

$$\|A\|_{\mathsf{op}} = O\left(\frac{1}{\sqrt{d}}\right), \tag{F.205}$$

with probability at least $1 - \exp(-c \log^2 n) - \exp(-cd)$ over V and X, where c is an absolute constant.

Proof. Recalling that $\nabla_z \phi(Vz)^\top = V^\top \mathsf{diag}\left(\phi'(Vz)\right)$, we have

$$\begin{aligned} \|A\|_{\mathsf{op}} &= \left\| \nabla_z \phi(Vz)^\top \phi(XV^\top)^\top K^{-1} \right\|_{\mathsf{op}} \\ &= \left\| V^\top \mathsf{diag}\left(\phi'(Vz)\right) \phi(XV^\top)^\top K^{-1} \right\|_{\mathsf{op}} \\ &\leq \|V\|_{\mathsf{op}} \left\| \mathsf{diag}\left(\phi'(Vz)\right) \right\|_{\mathsf{op}} \left\| \Phi^+ \right\|_{\mathsf{op}} \\ &\leq \|V\|_{\mathsf{op}} L \lambda_{\min}(K)^{-1/2}. \end{aligned} \tag{F.206}$$

where in the last line we use that ϕ is an L-Lipschitz function and that $K^{-1} = (\Phi^+)^\top(\Phi^+)$. By Theorem 4.4.5 of Vershynin [2018], we have (as $d = o(k)$ by Assumption 5.3) that

$$\|V\|_{\mathsf{op}} = O\left(1 + \sqrt{\frac{k}{d}}\right) = O\left(\sqrt{\frac{k}{d}}\right), \tag{F.207}$$

with probability at least $1 - \exp(-cd)$ over V. Lemma F.11 readily gives $\lambda_{\min}(K) = \Omega(k)$, with probability at least $1 - \exp(-c\log^2 n)$ over V and X. Taking a union bound on these two events gives the thesis. □

Proof of Theorem 5.5. By Theorem 5.8, we have $\|\mathcal{I}_{\mathsf{RF}}\|_F \geq C_1 \frac{k\sqrt{n}}{\sqrt{d}}$, with probability at least $1 - \exp(-c\log^2 k)$, which implies, by Theorem 5.7,

$$\|A(z)\|_F \geq C_2 \frac{d}{kn} \frac{k\sqrt{n}}{\sqrt{d}} - C\sqrt{n+d}/d = C_2 \frac{\sqrt{d}}{\sqrt{n}} - C\sqrt{n+d}/d \geq C_3 \frac{\sqrt{d}}{\sqrt{n}}, \tag{F.208}$$

where the last inequality follows from Assumption 5.4, and the first holds with probability at least $1 - \exp(-c\log^2 n) - k\exp(-cd)$. This also means that

$$\mathcal{S}(z) \geq C_4 \sqrt{d}\, \|A(z)\|_F \tag{F.209}$$

holds with probability at least $1 - \exp(-cd^2/n)$, because of Theorem 5.6 and Lemma F.27. Thus, we can conclude that

$$\mathcal{S}(z) \geq C_4 \|z\| \|A(z)\|_F \geq C_5 \frac{d}{\sqrt{n}} \geq C_6 \sqrt[6]{n}, \tag{F.210}$$

where the last inequality is a consequence of Assumption 5.4. □

F.3 Proofs for NTK Regression

In this section, we will indicate with $X \in \mathbb{R}^{n\times d}$ the data matrix, such that its rows are sampled independently from P_X (see Assumption 5.1). We will indicate with $W \in \mathbb{R}^{k\times d}$ the first half of the weight matrix at initialization, such that $W_{ij} \sim_{\text{i.i.d.}} \mathcal{N}(0, 1/d)$. The feature vector will be therefore be (see (5.26))

$$\Phi_{\mathsf{NTK}}(x) = [x \otimes \phi'(Wx), -x \otimes \phi'(Wx)] \tag{F.211}$$

where ϕ is the activation function, applied component-wise to the pre-activations Wx.

Notice that, for compactness, we dropped the subscripts "NTK" from these quantities, as this section will only treat the proofs related to Section 5.5. Always for the sake of compactness, we will not re-introduce such quantities in the statements or the proofs of the following Lemmas.

Lemma F.28. *We have that*

$$\lambda_{\min}(K) = \Omega(kd), \tag{F.212}$$

with probability at least $1 - ne^{-c\log^2 k} - e^{-c\log^2 n}$ *over* X *and* W.

Proof. The thesis follows from Theorem 3.1 of Bombari et al. [2022b]. Notice that our assumptions on the data distribution P_X are stronger, and that our initialization of the very last layer (which differs from their Gaussian initialization) does not change the result. Our assumption $k = O(d)$ respects the *loose pyramidal topology* condition, and we left the overparameterization assumption unchanged. A main difference is that we do not assume the activation function ϕ to be Lipschitz anymore. This, however, stop being a necessary assumption since we are working with a 2-layer neural network, and ϕ doesn't appear in the expression of Neural Tangent Kernel. □

Lemma F.29. *We have that*

$$\max_i \left|\phi'(Wx_i)^\top \phi'(Wz)\right| = O\left(\sqrt{k}\log k\right), \tag{F.213}$$

with probability at least $1 - \exp(-c\log^2 k)$ *over* X *and* W.

Proof. Let's look at the term $i = 1$, and let's call $x := x_1$ for compactness. As in Lemma F.26, we define z' as follows

$$z' = \frac{\sqrt{d}\left(I - \frac{xx^\top}{d}\right)z}{\left\|\left(I - \frac{xx^\top}{d}\right)z\right\|_2}. \tag{F.214}$$

We have $x^\top z' = 0$, and $\|z'\|_2 = \sqrt{d}$. Furthermore, by (F.186), we also have that $\|z - z'\|_2$ is a sub-Gaussian random variable, in the probability space of x. It will also be convenient to characterize $\|\phi'(Wz')\|_2$. We have, indicating with w_j the j-th row of W,

$$\begin{aligned}\|\phi'(Wz')\|_2^2 &= \sum_{j=1}^k \phi'^2(w_j^\top z') = \sum_{j=1}^k \mathbb{E}_W\left[\phi'^2(w_j^\top z')\right] + \sum_{j=1}^k \left(\phi'^2(w_j^\top z') - \mathbb{E}_W\left[\phi'^2(w_j^\top z')\right]\right) \\ &= k\mathbb{E}_\rho\left[\phi'^2(\rho)\right] + \sum_{j=1}^k \left(\phi'^2(\rho_j) - \mathbb{E}_\rho\left[\phi'^2(\rho)\right]\right),\end{aligned} \tag{F.215}$$

where the ρ and ρ_j's indicate independent standard Gaussian random variables. Since ϕ' is Lipschitz and non-zero, we have that $\mathbb{E}_\rho[\phi'^2(\rho)] = \Theta(1)$. Furthermore, by Bernstein inequality (cf. Theorem 2.8.1 in Vershynin [2018]), we have $\sum_{j=1}^k (\phi'^2(\rho_j) - \mathbb{E}_\rho[\phi'^2(\rho)]) = O(k)$ with probability at least $1 - 2\exp(-ck)$. This means that, with this probability (which is over W), we have

$$\|\phi'(Wz')\|_2 \le C\sqrt{k}, \tag{F.216}$$

where C is a numerical constant. Notice that the same result holds for $\|\phi'(Wx)\|_2$. We also have $\|W\|_{\text{op}} = O(1)$, which happens with probability at least $1 - \exp(-ck)$ over W, because of Theorem 4.4.5 of Vershynin [2018] and Assumption 5.11.

Since ϕ' is Lipschitz, we can write

$$\left|\phi'(Wx)^\top \phi'(Wz)\right| \le \left|\phi'(Wx)^\top \phi'(Wz')\right| + L\|\phi'(Wx)\|_2 \|W\|_{\text{op}} \|z - z'\|_2. \tag{F.217}$$

Let's look at the first term of the LHS. Since $x^\top z' = 0$, and $\|x\|_2 = \|z'\|_2 = \sqrt{d}$, we have

$$\phi'(Wx)^\top \phi'(Wz') = \phi'(\hat{\rho}_a)^\top \phi'(\hat{\rho}_b), \tag{F.218}$$

where $\hat{\rho}_a$ and $\hat{\rho}_b$ are two independent Gaussian vectors, each of them with independent entries. As ϕ is an even function (see Assumption 5.9), we have that $\phi'(\hat{\rho}_a)$ and $\phi'(\hat{\rho}_b)$ are independent sub-Gaussian vectors, with norm upper bounded by a numerical constant. Thus,

$$\mathbb{P}\left(\left|\phi'(Wx)^\top \phi'(Wz')\right| > C\sqrt{k}\log k\right) < 2\exp(-c_1 \log^2 k), \tag{F.219}$$

where this holds due to (F.216). Let's now look at the second term of the LHS of (F.217). We have

$$\|\phi'(Wx)\|_2 \|W\|_{\mathsf{op}} \|z - z'\|_2 \leq C\sqrt{k}C_1 \|z - z'\|_2 = O\left(\log k\sqrt{k}\right), \tag{F.220}$$

where the inequality holds with probability at least $1 - \exp(-ck)$ over W, and the equality holds with probability at least $1 - 2\exp(-c_2 \log^2 k)$ over x. Taking the intersection between the previously described high probability events, from (F.217) we get

$$\left|\phi'(Wx)^\top \phi'(Wz)\right| = O\left(\sqrt{k}\log k\right), \tag{F.221}$$

with probability at least $1 - 2\exp(-c\log^2 k)$ over x and W. Thus, performing a union bound over the different x_i's, we get the thesis. □

Lemma F.30.

$$\left\|\nabla_z \Phi(z)^\top \Phi(X)^\top\right\|_{\mathsf{op}} = O\left(\log k\left(\sqrt{k} + \sqrt{n}\right)\sqrt{d}\right), \tag{F.222}$$

with probability at least $1 - \exp(-c\log^2 k)$ *over* X, W *and* z.

Proof. We have, by application of the chain rule,

$$\begin{aligned}
\nabla_z \Phi(z)^\top \Phi(X)^\top &= 2\,\nabla_z \Phi'(z)^\top \Phi'(X)^\top \\
&= 2\,\nabla_z \left(z \otimes \phi'(Wz)\right)^\top \left(X * \phi'(XW^\top)\right)^\top \\
&= 2\left(I * \mathbf{1}\phi'^\top(Wz) + \mathbf{1}z^\top * W^\top \mathsf{diag}\left(\phi''(Wz)\right)\right)\left(X * \phi'(XW^\top)\right)^\top \\
&= 2\,X^\top \circ \mathbf{1}\left(\phi'(XW^\top)\phi'(Wz)\right)^\top + 2\,\mathbf{1}\left(Xz\right)^\top \circ W^\top \mathsf{diag}\left(\phi''(Wz)\right)\phi'(XW^\top)^\top \\
&= 2\,X^\top \mathsf{diag}\left(\phi'(XW^\top)\phi'(Wz)\right) + 2\,W^\top \mathsf{diag}\left(\phi''(Wz)\right)\phi'(XW^\top)^\top \mathsf{diag}\left(Xz\right).
\end{aligned} \tag{F.223}$$

Let's now condition on the event $\|W\|_{\mathsf{op}} = O(1)$, which happens with probability at least $1 - \exp(-cd)$ over W, because of Theorem 4.4.5 of Vershynin [2018] and Assumption 5.11.

Let's now look at the two terms separately. For the first, we have

$$\begin{aligned}
\left\|X^\top \mathsf{diag}\left(\phi'(XW^\top)\phi'(Wz)\right)\right\|_{\mathsf{op}} &\leq \|X\|_{\mathsf{op}} \max_i \left|\phi'(Wx_i)^\top \phi'(Wz)\right| \\
&= O\left(\log k\sqrt{k}\left(\sqrt{n} + \sqrt{d}\right)\right),
\end{aligned} \tag{F.224}$$

where the last equality is true because of Lemmas F.5 and F.29 and holds with probability at least $1 - \exp(-cd) - \exp(-c\log^2 k)$ over X and W. For the other term, we have

$$\begin{aligned}
\left\|W^\top \mathsf{diag}\left(\phi''(Wz)\right)\phi'(XW^\top)\mathsf{diag}\left(Xz\right)\right\|_{\mathsf{op}} &\leq \|W\|_{\mathsf{op}} L \left\|\phi'(XW^\top)\right\|_{\mathsf{op}} \max_i \left|x_i^\top z\right| \\
&= O\left(\log k\sqrt{d}\left(\sqrt{n} + \sqrt{k}\right)\right).
\end{aligned} \tag{F.225}$$

The last step uses Lemma F.5 and the fact that we are taking the maximum over k sub-Gaussian random variables such that $\left\|x_i^\top z\right\|_{\psi_2} \leq C\|z\|_2 = C\sqrt{d}$, and it holds with probability at least $1 - 2\exp(-ck) - \exp(-c\log^2 k)$ over X.

Putting everything together, and using Assumption 5.11, we get

$$\begin{aligned}\left\|\nabla_z \Phi(z)^\top \Phi(X)^\top\right\|_{\text{op}} &= O\left(\log k\sqrt{k}\left(\sqrt{n}+\sqrt{d}\right)\right) + O\left(\log k\sqrt{d}\left(\sqrt{n}+\sqrt{k}\right)\right) \\ &= O\left(\log k\left(\sqrt{k}+\sqrt{n}\right)\sqrt{d}\right),\end{aligned} \tag{F.226}$$

which gives the thesis. □

F.3.1 Proof of Theorem 5.12

Proof of Theorem 5.12. We have

$$\begin{aligned}\mathcal{S}(z) &\leq \sqrt{d}\left\|\nabla_z \Phi(z)^\top \Phi(X)^\top\right\|_{\text{op}} \left\|K^{-1}\right\|_{\text{op}} \|Y\|_2 \\ &= \sqrt{d} O\left(\log k\left(\sqrt{k}+\sqrt{n}\right)\sqrt{d}\right) O\left(\frac{1}{dk}\right) O\left(\sqrt{n}\right) \\ &= O\left(\log k\left(1+\sqrt{n/k}\right)\frac{\sqrt{n}}{\sqrt{k}}\right),\end{aligned} \tag{F.227}$$

where the second line is justified by Lemmas F.28 and F.30, and holds with probability at least $1 - ne^{-c\log^2 k} - e^{-c\log^2 n}$ over X and W. The second statement is immediate. □

F.4 Additional Experiments

In this section, we provide additional experimental results for the activation functions $\phi(x) = \cos(x)$ (which is even) and $\phi(x) = \frac{e^x}{1+e^x}$ (namely, softplus or smooth ReLU, which is not even). Again, our theoretical findings are supported by numerical evidence both on the RF (see Figure F.1) and the NTK regression model (see Figure F.2).

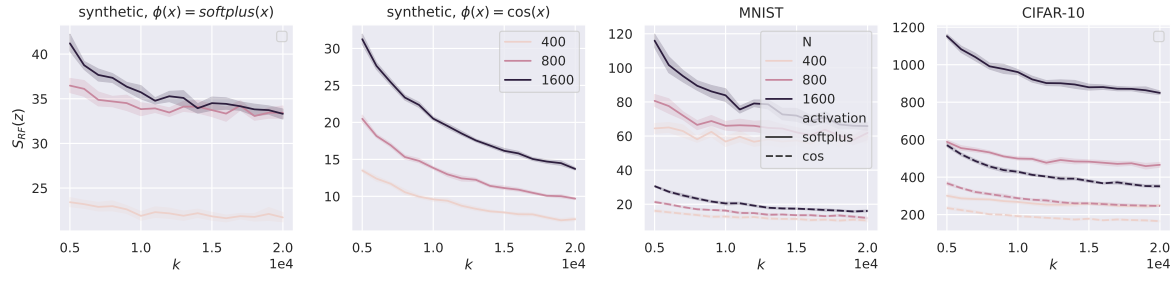


Figure F.1: Sensitivity as a function of the number of parameters k, for an RF model with synthetic data sampled from a Gaussian distribution with input dimension $d = 1000$ (first and second plot), and with inputs taken from two classes of the MNIST and CIFAR-10 datasets (third and fourth plot respectively). Different curves correspond to different values of the sample size $n \in \{400, 800, 1600\}$ and to different activations ($\phi(x) = \text{softplus}(x)$ or $\phi(x) = \cos(x)$). We plot the average over 10 independent trials and the confidence interval at 1 standard deviation. The values of the sensitivity are significantly larger for $\phi(x) = \text{softplus}(x)$ than for $\phi(x) = \cos(x)$.

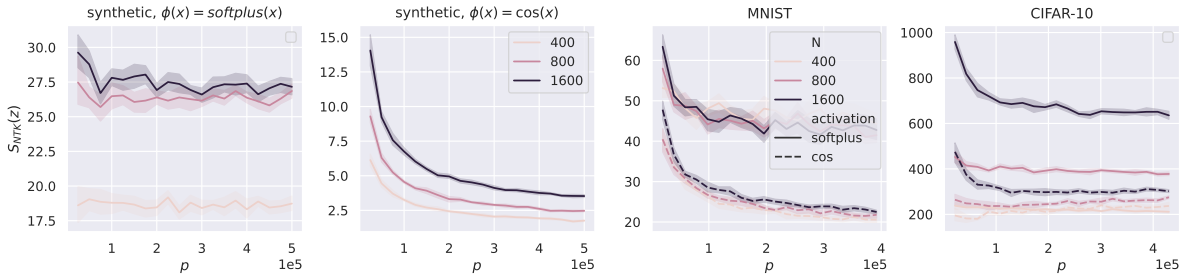


Figure F.2: Sensitivity as a function of the number of parameters p, for an NTK model with synthetic data sampled from a Gaussian distribution with input dimension $d = 1000$ (first and second plot), and with inputs taken from two classes of the MNIST and CIFAR-10 datasets (third and fourth plot respectively). Different curves correspond to different values of the sample size $n \in \{400, 800, 1600\}$ and to different activations ($\phi(x) = \text{softplus}(x)$ or $\phi(x) = \cos(x)$). We plot the average over 10 independent trials and the confidence interval at 1 standard deviation. The values of the sensitivity are significantly larger for $\phi(x) = \text{softplus}(x)$ than for $\phi(x) = \cos(x)$.

APPENDIX G

Appendix for Chapter 6

G.1 Proofs for Linear Regression

Proof of Proposition 6.2. Note that

$$\hat{\theta}_{\mathsf{LR}}(0) = \left(Z^\top Z\right)^{-1} Z^\top G. \tag{G.1}$$

Since we have $g_i = z_i^\top \theta^* + \epsilon_i$, (G.1) reads

$$\hat{\theta}_{\mathsf{LR}}(0) = \left(Z^\top Z\right)^{-1} Z^\top \left(Z\theta^* + \mathcal{E}\right) = \theta^* + \left(Z^\top Z\right)^{-1} Z^\top \mathcal{E}. \tag{G.2}$$

Then, we can plug this result in the definition of $\mathcal{C}(\hat{\theta})$ in (6.3) to obtain

$$\begin{aligned}
\mathbb{E}_{\mathcal{E}}\left[\mathcal{C}(\hat{\theta}_{\mathsf{LR}}(0))\right] &= \mathbb{E}_{\mathcal{E}}\left[\mathrm{Cov}_{[x^\top, y^\top]^\top \sim P_{XY},\, g = f_x^*(x),\, \tilde{x}\sim P_X}\left(f_{\mathsf{LR}}\left(\hat{\theta}_{\mathsf{LR}}(0), [\tilde{x}^\top, y^\top]^\top\right), g\right)\right] \\
&= \mathbb{E}_{\mathcal{E}}\left[\mathrm{Cov}_{[x^\top, y^\top]^\top \sim P_{XY},\, \tilde{x}\sim P_X}\left([\tilde{x}^\top, y^\top]\hat{\theta}_{\mathsf{LR}}(0), x^\top \theta_x^*\right)\right] \\
&= \mathbb{E}_{\mathcal{E}}\left[\mathrm{Cov}_{[x,y]\sim P_{XY},\, \tilde{x}\sim P_X}\left(\tilde{x}^\top\theta_x^* + [\tilde{x}^\top, y^\top]\left(Z^\top Z\right)^{-1} Z^\top \mathcal{E}, x^\top \theta_x^*\right)\right] \\
&= \mathrm{Cov}_{[x,y]\sim P_{XY},\, \tilde{x}\sim P_X}\left(\tilde{x}^\top\theta_x^*, x^\top\theta_x^*\right) \\
&= 0,
\end{aligned} \tag{G.3}$$

where in the second line we used that $\mathcal{E}$ is independent from everything else, in fourth line we used $\mathbb{E}[\mathcal{E}] = 0$, and that $\mathcal{E}$ is independent from all the other random variables, and the last step holds since $\tilde{x}$ is independent from x.

For the second part of the statement we have that

$$\begin{aligned}
\mathcal{C}(\hat{\theta}_{\mathsf{LR}}(0)) &= \mathrm{Cov}_{[x^\top, y^\top]^\top \sim P_{XY},\, \tilde{x}\sim P_X}\left([\tilde{x}^\top, y^\top]\left(Z^\top Z\right)^{-1} Z^\top \mathcal{E}, x^\top\theta_x^*\right) \\
&= \mathrm{Cov}_{[x^\top, y^\top]^\top \sim P_{XY},\, \tilde{x}\sim P_X}\left(\mathcal{E}^\top Z\left(Z^\top Z\right)^{-1} P_y [x^\top, y^\top]^\top, [x^\top, y^\top]\theta^*\right) \\
&= \mathcal{E}^\top Z\left(Z^\top Z\right)^{-1} P_y \Sigma \theta^*,
\end{aligned} \tag{G.4}$$

where in the second line we introduced $P_y \in \mathbb{R}^{2d\times 2d}$, defined as the projector on the last d elements of the canonical basis in $\mathbb{R}^{2d}$. Then, since $\mathcal{E}$ is a sub-Gaussian vector (the entries are

mean-0, i.i.d. sub-Gaussian) independent from everything else, we have that, with probability at least $1 - 2\exp\left(-c_1 \log^2 d\right)$,

$$\begin{aligned}\left|\mathcal{C}(\hat{\theta}_{\mathsf{LR}}(0))\right| &\leq \log d \left\| Z\left(Z^\top Z\right)^{-1} P_y \Sigma \theta^* \right\|_2 \\ &\leq \log d \left\| Z\left(Z^\top Z\right)^{-1} \right\|_{\mathsf{op}} \|P_y\|_{\mathsf{op}} \|\Sigma\|_{\mathsf{op}} \|\theta^*\|_2 \\ &\leq \frac{\log d \|\Sigma\|_{\mathsf{op}}}{\sqrt{\lambda_{\min}\left(Z^\top Z\right)}},\end{aligned} \tag{G.5}$$

where we used $\|P_y\|_{\mathsf{op}} = 1$ and $\|\theta^*\|_2 = 1$. Since Z is a $n \times 2d$ matrix with independent rows having second moment Σ, by Theorem 5.39 in Vershynin [2012] (see Remark 5.40), we have that

$$\left\| \frac{Z^\top Z}{n} - \Sigma \right\|_{\mathsf{op}} = O\left(\sqrt{\frac{d}{n}}\right) = o(1), \tag{G.6}$$

with probability at least $1 - 2\exp\left(-c_2 d\right)$. Hence, with this probability, by Weyl's inequality, we also have

$$\lambda_{\min}\left(Z^\top Z\right) \geq n\lambda_{\min}\left(\Sigma\right) - \left\| Z^\top Z - n\Sigma \right\|_{\mathsf{op}} = \Theta(n), \tag{G.7}$$

where the last step holds because of Assumption 6.1. Thus, we have that (G.5) reads

$$\left|\mathcal{C}(\hat{\theta}_{\mathsf{LR}}(0))\right| \leq \frac{\log d \|\Sigma\|_{\mathsf{op}}}{\sqrt{\lambda_{\min}\left(Z^\top Z\right)}} = O\left(\frac{\log d}{\sqrt{n}}\right), \tag{G.8}$$

with probability at least $1 - 2\exp\left(-c_3 \log^2 d\right)$ over Z and $\mathcal{E}$, which gives the desired result. □

Proof of Theorem 6.3. As in Han and Xu [2023], we define the Gaussian sequence model $\hat{\theta}^\rho \in \mathbb{R}^{2d}$ as

$$\hat{\theta}^\rho = \left(\Sigma + \tau(\lambda) I\right)^{-1} \Sigma^{1/2} \left(\Sigma^{1/2}\theta^* + \frac{\gamma\rho}{\sqrt{2d}}\right), \tag{G.9}$$

where ρ is a standard Gaussian vector in $\mathbb{R}^{2d}$ (the Gaussian sequence model is defined in Equation (1.5) in Han and Xu [2023], via the different notation $\hat{\mu}^{\mathsf{seq}}_{(\Sigma,\mu_0)}$, and our Gaussian vector ρ is denoted as g in the same equation). In the equation above, $\gamma > 0$ is implicitly defined via

$$\frac{n\gamma^2}{2d} = \sigma^2 + \mathbb{E}_\rho\left[\left\|\Sigma^{1/2}\left(\hat{\theta}^\rho - \theta^*\right)\right\|_2^2\right]. \tag{G.10}$$

On the other hand, following a similar argument as the one in (G.3), we have that, for every $\theta \in \mathbb{R}^{2d}$,

$$\begin{aligned}\mathcal{C}(\theta) &= \mathrm{Cov}_{[x^\top, y^\top]^\top \sim P_{XY},\, \tilde{x} \sim P_X}\left([\tilde{x}^\top, y^\top]\theta, x^\top \theta^*_x\right) \\ &= \theta^\top \mathbb{E}_{[x^\top, y^\top]^\top \sim P_{XY},\, \tilde{x} \sim P_X}\left[[\tilde{x}^\top, y^\top]^\top x^\top\right] \theta^*_x \\ &= \theta^\top \mathbb{E}_{[x^\top, y^\top]^\top \sim P_{XY}}\left[[\mathbf{0}^\top, y^\top]^\top [x^\top, \mathbf{0}^\top]\right] \theta^* \\ &= \theta^\top P_y \mathbb{E}_{[x^\top, y^\top]^\top \sim P_{XY}}\left[[x^\top, y^\top]^\top [x^\top, y^\top]\right] \theta^* \\ &= \theta^\top P_y \Sigma \theta^*,\end{aligned} \tag{G.11}$$

where the third line holds since $\tilde{x}$ has 0 mean and is independent from x and y, and by definition of θ^*_x, and the fourth line holds because $P_y[x^\top, y^\top]^\top = [\mathbf{0}^\top, y^\top]^\top$ and because the last d entries

of θ^* are 0 (*i.e.*, $P_y\theta^* = 0$). Thus, since we have that $\|P_y\Sigma\theta^*\|_2 \leq \|P_y\|_{\mathsf{op}} \|\Sigma\|_{\mathsf{op}} \|\theta^*\|_2 \leq \|\Sigma\|_{\mathsf{op}}$ because of Assumption 6.1 (and since $\|\theta^*\| \leq 1$), we have that $\mathcal{C}(\cdot) : \mathbb{R}^{2d} \to \mathbb{R}$ is a $\|\Sigma\|_{\mathsf{op}}$-Lipschitz function.

Now, since $\mathcal{P}_{XY}$ is multivariate Gaussian, Theorem 2.3 of Han and Xu [2023] gives that, for any 1-Lipschitz function $\varphi : \mathbb{R}^{2d} \to \mathbb{R}$ (denoted as g in Han and Xu [2023]), and any $t \in (0, 1/2)$,

$$\mathbb{P}_{Z,G}\left(\left|\varphi(\hat{\theta}_{\mathsf{LR}}(\lambda)) - \mathbb{E}_\rho\left[\varphi\left(\hat{\theta}^\rho\right)\right]\right| \geq t\right) \leq C_1 d \exp\left(-dt^4/C_1\right), \tag{G.12}$$

where C_1 is a constant depending on $\lambda_{\min}(\Sigma)$, $\|\Sigma\|_{\mathsf{op}}$, σ^2, and $n/d = \Theta(1)$. Since $\mathcal{C}(\cdot)$ is linear, notice that we have

$$\mathbb{E}_\rho\left[\mathcal{C}\left(\hat{\theta}^\rho\right)\right] = \mathcal{C}\left(\mathbb{E}_\rho\left[\hat{\theta}^\rho\right]\right) = \mathcal{C}\left(\left(\Sigma + \tau(\lambda)I\right)^{-1}\Sigma\theta^*\right) = \theta^{*\top}\Sigma\left(\Sigma + \tau(\lambda)I\right)^{-1} P_y\Sigma\theta^* = \mathcal{C}^\Sigma(\lambda), \tag{G.13}$$

where we used (G.11) in the third step, and the definition of $\mathcal{C}^\Sigma(\lambda)$ in (6.12) in the last one. Thus, setting $\varphi(\cdot)$ to be $\mathcal{C}(\cdot)/\|\Sigma\|_{\mathsf{op}}$, and plugging (G.13) in (G.12) we obtain

$$\mathbb{P}_{Z,G}\left(\frac{\left|\mathcal{C}(\hat{\theta}_{\mathsf{LR}}(\lambda)) - \mathcal{C}^\Sigma(\lambda)\right|}{\|\Sigma\|_{\mathsf{op}}} \geq t\right) \leq C_1 d \exp\left(-dt^4/C_1\right), \tag{G.14}$$

which gives the thesis after absorbing the constant $\|\Sigma\|_{\mathsf{op}}$ in t, and noticing that the bound is still true for $t \in (0, 1/2)$ since $\|\Sigma\|_{\mathsf{op}} \geq \mathrm{tr}(\Sigma)/2d = 1$ by Assumption 6.1. □

Proposition G.1. *Let $\mathcal{C}^\Sigma(\lambda)$ be defined in* (6.12)*, and let $S_x^{\Sigma+\tau(\lambda)I}$ be the Schur complement of $\Sigma + \tau(\lambda)I$ with respect to the top-left $d \times d$ block. Then, we have that*

$$\mathcal{C}^\Sigma(\lambda) = \tau(\lambda)\,\theta_x^{*\top}\left(\Sigma_{xx} + \tau(\lambda)I\right)^{-1}\Sigma_{xy}\left(S_x^{\Sigma+\tau(\lambda)I}\right)^{-1}\Sigma_{yx}\theta_x^*. \tag{G.15}$$

Proof. During the proof, to ease the notation, we will often leave implicit the dependence of τ on λ. Then, we can write

$$\begin{aligned}
\mathcal{C}^\Sigma(\lambda) &= \theta^{*\top}\Sigma\left(\Sigma + \tau I\right)^{-1} P_y\Sigma\theta^* \\
&= \theta^{*\top}(\Sigma + \tau I - \tau I)\left(\Sigma + \tau I\right)^{-1} P_y\Sigma\theta^* \\
&= -\tau\theta^{*\top}\left(\Sigma + \tau I\right)^{-1} P_y\Sigma\theta^* + \theta^{*\top}P_y\Sigma\theta^* \\
&= -\tau\theta^{*\top}\left(\Sigma + \tau I\right)^{-1} P_y\Sigma\theta^*,
\end{aligned} \tag{G.16}$$

where the last step holds since $P_y\theta^* = 0$. This expression can be further manipulated using the notation introduced in (6.14). We also introduce the following notation

$$\left(\Sigma + \tau I\right)^{-1} = \left(\begin{array}{c|c} \left[\left(\Sigma + \tau I\right)^{-1}\right]_{xx} & \left[\left(\Sigma + \tau I\right)^{-1}\right]_{xy} \\ \hline \left[\left(\Sigma + \tau I\right)^{-1}\right]_{yx} & \left[\left(\Sigma + \tau I\right)^{-1}\right]_{yy} \end{array}\right), \tag{G.17}$$

where we divided $\left(\Sigma + \tau I\right)^{-1}$ in four $d \times d$ blocks. Notice that, the expression in (G.16) only depends on $\left[\left(\Sigma + \tau I\right)^{-1}\right]_{xy}$, *i.e.*,

$$\mathcal{C}^\Sigma(\lambda) = -\tau\theta^{*\top}P_x\left(\Sigma + \tau I\right)^{-1} P_y\Sigma\theta^* = -\tau\theta_x^{*\top}\left[\left(\Sigma + \tau I\right)^{-1}\right]_{xy}\Sigma_{yx}\theta_x^*, \tag{G.18}$$

where we denoted the projector on the first d elements of the canonical basis of $\mathbb{R}^{2d}$ as $P_x \in \mathbb{R}^{2d\times 2d}$. Exploiting the Schur complement $S_x^{\Sigma+\tau I}$, it holds that

$$\left[\left(\Sigma + \tau I\right)^{-1}\right]_{xy} = -\left(\Sigma_{xx} + \tau I\right)^{-1}\Sigma_{xy}\left(S_x^{\Sigma+\tau I}\right)^{-1}, \tag{G.19}$$

which combined with (G.18) proves (G.15). □

Proof of Proposition 6.4. During the proof, to ease the notation, we will often leave implicit the dependence of τ on λ. Then, according to (6.12), we have that

$$\left|\mathcal{C}^{\Sigma}(\lambda)\right| = \left|\theta^{*\top}\Sigma\left(\Sigma+\tau I\right)^{-1}P_y\Sigma P_x\theta^*\right| \leq \|\theta^*\|_2^2 \left\|\Sigma\left(\Sigma+\tau I\right)^{-1}\right\|_{\mathsf{op}} \|P_y\Sigma P_x\|_{\mathsf{op}} \leq \|\Sigma_{yx}\|_{\mathsf{op}}, \tag{G.20}$$

and

$$\left|\mathcal{C}^{\Sigma}(\lambda)\right| = \left|\theta^{*\top}\Sigma\left(\Sigma+\tau I\right)^{-1}P_y\Sigma\theta^*\right| \leq \|\theta^*\|_2^2 \|\Sigma\|_{\mathsf{op}}^2 \frac{1}{\lambda_{\min}\left(\Sigma\right)+\tau} \leq \frac{\lambda_{\max}\left(\Sigma\right)^2}{\tau}. \tag{G.21}$$

Then, using (G.15), we get

$$\begin{aligned}
&\mathcal{C}^{\Sigma}(\lambda) = \\
&= \tau\,\theta_x^{*\top}\left(\Sigma_{xx}+\tau I\right)^{-1}\Sigma_{xy}\left(S_x^{\Sigma+\tau I}\right)^{-1}\Sigma_{yx}\theta_x^* \\
&= \tau\,\theta_x^{*\top}\left(\Sigma_{xx}+\tau I\right)^{-1/2}\left(\Sigma_{xx}+\tau I\right)^{-1/2}\Sigma_{xy}\left(S_x^{\Sigma+\tau I}\right)^{-1}\Sigma_{yx}\left(\Sigma_{xx}+\tau I\right)^{-1/2}\left(\Sigma_{xx}+\tau I\right)^{1/2}\theta_x^* \\
&\leq \tau\,\frac{1}{\sqrt{\lambda_{\min}\left(\Sigma_{xx}\right)+\tau}}\sqrt{\theta_x^{*\top}\Sigma_{xx}\theta_x^*+\tau}\left\|\left(\Sigma_{xx}+\tau I\right)^{-1/2}\Sigma_{xy}\left(S_x^{\Sigma+\tau I}\right)^{-1}\Sigma_{yx}\left(\Sigma_{xx}+\tau I\right)^{-1/2}\right\|_{\mathsf{op}} \\
&\leq \tau\frac{\sqrt{\mathbb{E}_{x\sim\mathcal{P}_X}\left[\left(x^\top\theta_x^*\right)^2\right]+\tau}}{\sqrt{\lambda_{\min}\left(\Sigma_{xx}\right)+\tau}}\frac{\left\|\left(\Sigma_{xx}+\tau I\right)^{-1/2}\Sigma_{xy}\right\|_{\mathsf{op}}^2}{\lambda_{\min}\left(S_x^{\Sigma+\tau I}\right)} \\
&= \tau\frac{\sqrt{\mathbb{E}_{g=x^\top\theta_x^*+\epsilon}\left[g^2\right]-\sigma^2+\tau}}{\sqrt{\lambda_{\min}\left(\Sigma_{xx}\right)+\tau}}\frac{\lambda_{\max}\left(\Sigma_{yx}\left(\Sigma_{xx}+\tau I\right)^{-1}\Sigma_{xy}\right)}{\lambda_{\min}\left(\Sigma_{yy}+\tau I-\Sigma_{yx}\left(\Sigma_{xx}+\tau I\right)^{-1}\Sigma_{xy}\right)} \\
&\leq \tau\frac{\sqrt{\mathbb{E}_g\left[g^2\right]-\sigma^2+\tau}}{\sqrt{\lambda_{\min}\left(\Sigma_{xx}\right)+\tau}}\frac{\lambda_{\max}\left(\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}\right)}{\lambda_{\min}\left(\Sigma_{yy}+\tau I-\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}\right)} \\
&= \tau\frac{\sqrt{\mathbb{E}_g\left[g^2\right]-\sigma^2+\tau}}{\sqrt{\lambda_{\min}\left(\Sigma_{xx}\right)+\tau}}\frac{\lambda_{\max}\left(\Sigma_{yy}\right)-\lambda_{\min}\left(S_x^{\Sigma}\right)}{\lambda_{\min}\left(S_x^{\Sigma}\right)+\tau} \\
&\leq \tau\sqrt{\mathrm{Var}(g)-\sigma^2}\frac{\lambda_{\max}\left(\Sigma_{yy}\right)-\lambda_{\min}\left(S_x^{\Sigma}\right)}{\lambda_{\min}\left(S_x^{\Sigma}\right)\sqrt{\lambda_{\min}\left(\Sigma_{xx}\right)}},
\end{aligned} \tag{G.22}$$

where in the fifth line we denoted with $\mathcal{P}_X$ the marginal distribution of the core feature x, and $\mathbb{E}_g\left[\cdot\right]$ from the sixth line on denotes an expectation with respect to g distributed as the labels of the model. The last step simplifies the expression with respect to τ, and it holds since $\mathrm{Var}(g)-\sigma^2 = \theta_x^{*\top}\Sigma_{xx}\theta_x^* \geq \lambda_{\min}\left(\Sigma_{xx}\right)$. This, together with (G.20) and (G.21) gives the desired result. □

Proof of Proposition 6.5. During the proof, to ease the notation, we will often leave implicit the dependence of τ on λ. Then, as in Han and Xu [2023] and in the proof of Theorem 6.3, we define the Gaussian sequence model $\hat{\theta}^{\rho} \in \mathbb{R}^{2d}$ as (G.9) where ρ is a standard Gaussian vector in $\mathbb{R}^{2d}$ and $\gamma > 0$ is implicitly defined via

$$\begin{aligned}
\frac{n\gamma^2}{2d} &= \sigma^2 + \mathbb{E}_{\rho}\left[\left\|\Sigma^{1/2}\left(\hat{\theta}^{\rho}-\theta^*\right)\right\|_2^2\right] \\
&= \sigma^2 + \left\|\Sigma^{1/2}\left(\left(\Sigma+\tau I\right)^{-1}\Sigma - I\right)\theta^*\right\|_2^2 + \frac{\gamma^2}{2d}\,\mathrm{tr}\left(\left(\Sigma+\tau I\right)^{-2}\Sigma^2\right),
\end{aligned} \tag{G.23}$$

which also reads

$$\frac{n\gamma^2}{2d} = \frac{\sigma^2 + \tau^2 \left\| (\Sigma + \tau I)^{-1} \Sigma^{1/2} \theta^* \right\|_2^2}{1 - \frac{\mathrm{tr}\left((\Sigma+\tau I)^{-2}\Sigma^2\right)}{n}}. \tag{G.24}$$

Then, due to Theorem 3.1 of Han and Xu [2023] on the prediction risk, since $\mathcal{P}_{XY}$ is a multivariate Gaussian due to Assumption 6.1, we have that, for any $t \in (0, 1/2)$,

$$\mathbb{P}_{Z,G}\left(\left| \mathcal{L}(\hat{\theta}_{\mathsf{LR}}(\lambda)) - \mathcal{L}^{\Sigma}(\lambda) \right| \geq t \right) \leq Cd \exp\left(-dt^4/C \right), \tag{G.25}$$

where C is a positive constant depending on $\lambda_{\min}(\Sigma)$, $\|\Sigma\|_{\mathsf{op}}$, σ^2, and $n/d = \Theta(1)$. □

Proof of Proposition 6.6 As $\tau(\lambda)$ is an increasing function of λ, all the statements on the monotonicity of $\mathcal{L}^{\Sigma}(\lambda)$ and $\mathcal{C}^{\Sigma}(\lambda)$ can be proved by showing monotonicity w.r.t. τ (whose dependence w.r.t. λ is left implicit throughout the argument). In particular, we have

$$\begin{aligned} \frac{\mathsf{d}\mathcal{L}^{\Sigma}(\lambda)}{\mathsf{d}\tau} = & \frac{\frac{\mathsf{d}}{\mathsf{d}\tau}\left(\tau^2 \left\| (\Sigma + \tau I)^{-1} \Sigma^{1/2} \theta^* \right\|_2^2 \right) \left(1 - \frac{\mathrm{tr}\left((\Sigma+\tau I)^{-2}\Sigma^2\right)}{n} \right)}{\left(1 - \frac{\mathrm{tr}\left((\Sigma+\tau I)^{-2}\Sigma^2\right)}{n} \right)^2} \\ & - \frac{\left(\sigma^2 + \tau^2 \left\| (\Sigma + \tau I)^{-1} \Sigma^{1/2} \theta^* \right\|_2^2 \right) \frac{\mathsf{d}}{\mathsf{d}\tau}\left(1 - \frac{\mathrm{tr}\left((\Sigma+\tau I)^{-2}\Sigma^2\right)}{n} \right)}{\left(1 - \frac{\mathrm{tr}\left((\Sigma+\tau I)^{-2}\Sigma^2\right)}{n} \right)^2}. \end{aligned} \tag{G.26}$$

To study the sign of the above expression, it suffices to focus on the numerators, as the denominator is always positive.

Note that the RHS of (6.13) is smaller or equal to $2d/n$; thus, as $2d < n$, we also get $\tau \leq \lambda (1 - 2d/n)^{-1}$, which implies $\tau(\lambda) \to 0$ as $\lambda \to 0$. Hence, to show that $\mathcal{L}^{\Sigma}(\lambda)$ is monotonically decreasing in a right neighborhood of $\lambda = 0$, it suffices to show that (G.26) evaluated in $\tau = 0$ is strictly negative. For $\tau = 0$, the first factor in the numerator of the first term in (G.26) is 0, as the following chain of equalities holds:

$$\begin{aligned} \frac{\mathsf{d}}{\mathsf{d}\tau}\left(\tau^2 \left\| (\Sigma + \tau I)^{-1} \Sigma^{1/2} \theta^* \right\|_2^2 \right) &= \frac{\mathsf{d}}{\mathsf{d}\tau}\left(\left\| (\Sigma/\tau + I)^{-1} \Sigma^{1/2} \theta^* \right\|_2^2 \right) \\ &= {\theta^*}^{\top} \frac{\mathsf{d}}{\mathsf{d}\tau} (\Sigma/\tau + I)^{-2} \Sigma \theta^* \\ &= -{\theta^*}^{\top} (\Sigma/\tau + I)^{-2} \left(\frac{\mathsf{d}}{\mathsf{d}\tau} (\Sigma/\tau + I)^2 \right) (\Sigma/\tau + I)^{-2} \Sigma \theta^* \\ &= -{\theta^*}^{\top} (\Sigma/\tau + I)^{-2} \left(-\frac{2\Sigma}{\tau^2} (\Sigma/\tau + I) \right) (\Sigma/\tau + I)^{-2} \Sigma \theta^* \\ &= 2\tau {\theta^*}^{\top} (\Sigma + \tau I)^{-3} \Sigma^2 \theta^*. \end{aligned} \tag{G.27}$$

Furthermore, the second term gives

$$-\sigma^2 \frac{\mathsf{d}}{\mathsf{d}\tau}\left(1 - \frac{\mathrm{tr}\left((\Sigma + \tau I)^{-2} \Sigma^2 \right)}{n} \right) = \frac{\sigma^2}{n} \frac{\mathsf{d}}{\mathsf{d}\tau}\left(\sum_{k=1}^{2d} \frac{\lambda_k^2}{(\lambda_k + \tau)^2} \right) = -\frac{2\sigma^2}{n} \sum_{k=1}^{2d} \frac{\lambda_k^2}{(\lambda_k + \tau)^3} < 0, \tag{G.28}$$

where λ_k denotes the k-th eigenvalue of Σ. This gives the first claim.

To show that there exists $\lambda_{\mathcal{L}} > 0$ such that $\mathcal{L}^{\Sigma}(\lambda)$ is monotonically increasing for $\lambda \geq \lambda_{\mathcal{L}}$ we will show that the derivative of $\mathcal{L}^{\Sigma}(\lambda)$ with respect to τ is positive for all $\tau \geq \tau_{\mathcal{L}} := \tau(\lambda_{\mathcal{L}})$. For simplicity, in the rest of the argument we use the notation $\lambda_{\max}$ and $\lambda_{\min}$ to indicate the largest and smallest eigenvalues of Σ, respectively. Instead, the notation $\lambda_{\min}(\cdot)$ still represents the smallest eigenvalue of its argument. For the first factor of the first term of (G.26), continuing from (G.27), we have

$$\frac{\mathsf{d}}{\mathsf{d}\tau}\left(\tau^2 \left\|(\Sigma+\tau I)^{-1}\Sigma^{1/2}\theta^*\right\|_2^2\right) \geq 2\frac{1}{\lambda_{\max}\left(\Sigma/\tau+I\right)^3}\lambda_{\min}\left(\Sigma/\tau\right)^2 = \frac{2\lambda_{\min}^2}{\tau^2\left(\lambda_{\max}/\tau+1\right)^3}. \tag{G.29}$$

For the second factor of the first term of (G.26), we have

$$1-\frac{\operatorname{tr}\left(\left(\Sigma+\tau I\right)^{-2}\Sigma^2\right)}{n} = 1-\frac{1}{n}\sum_{k=1}^{2d}\frac{\lambda_k^2}{\left(\lambda_k+\tau\right)^2} \geq 1-\frac{2d\lambda_{\max}^2}{n\tau^2}. \tag{G.30}$$

For the first factor of the second term of (G.26), we have

$$\tau^2\left\|(\Sigma+\tau I)^{-1}\Sigma^{1/2}\theta^*\right\|_2^2 \leq \tau^2\frac{\lambda_{\max}}{\left(\lambda_{\min}+\tau\right)^2} \tag{G.31}$$

For the second factor of the second term of (G.26), we have

$$\frac{\mathsf{d}}{\mathsf{d}\tau}\left(1-\frac{\operatorname{tr}\left(\left(\Sigma+\tau I\right)^{-2}\Sigma^2\right)}{n}\right) = -\frac{1}{n}\frac{\mathsf{d}}{\mathsf{d}\tau}\left(\sum_{k=1}^{2d}\frac{\lambda_k^2}{\left(\lambda_k+\tau\right)^2}\right) = \frac{2}{n}\sum_{k=1}^{2d}\frac{\lambda_k^2}{\left(\lambda_k+\tau\right)^3} \leq \frac{4d\lambda_{\max}^2}{n\tau^3}. \tag{G.32}$$

Thus, putting together (G.26), (G.29), (G.30), (G.31), and (G.32), the monotonicity of $\mathcal{L}^{\Sigma}(\lambda)$ is implied by

$$\frac{2\lambda_{\min}^2}{\tau^2\left(\lambda_{\max}/\tau+1\right)^3}\left(1-\frac{2d\lambda_{\max}^2}{n\tau^2}\right) \overset{?}{\geq} \left(\sigma^2+\tau^2\frac{\lambda_{\max}}{\left(\lambda_{\min}+\tau\right)^2}\right)\frac{4d\lambda_{\max}^2}{n\tau^3}. \tag{G.33}$$

Now, we have that the above inequality holds for sufficiently large τ: the LHS is $\Theta(1/\tau^2)$ (considering fixed the other quantities), while the RHS is $\Theta(1/\tau^3)$; and the desired statement is therefore proved.

Next, we set $\tau_{\mathcal{C}} := \tau(\lambda_{\mathcal{C}}) = \sqrt{\lambda_{\min}\left(S_x^{\Sigma}\right)}$ and show that $\mathcal{C}^{\Sigma}(\lambda)$ is monotonically increasing for $\tau \in [0, \tau_{\mathcal{C}}]$. Plugging $\Sigma_{xx} = I$ in (G.15) we get

$$\begin{aligned}\mathcal{C}^{\Sigma}(\lambda) &= \tau\,\theta_x^{*\top}\left(\Sigma_x+\tau I\right)^{-1}\Sigma_{xy}\left(S_x^{\Sigma+\tau I}\right)^{-1}\Sigma_{yx}\theta_x^* \\ &= \frac{\tau}{1+\tau}\,\theta_x^{*\top}\Sigma_{xy}\left(\Sigma_{yy}+\tau I-\frac{\Sigma_{yx}\Sigma_{xy}}{1+\tau}\right)^{-1}\Sigma_{yx}\theta_x^*.\end{aligned} \tag{G.34}$$

By the product rule, and introducing the shorthand $A(\tau) = \Sigma_{yy} + \tau I - \frac{\Sigma_{xy}^\top \Sigma_{xy}}{1+\tau}$, we have

$$
\begin{aligned}
&\frac{\mathsf{d}C^\Sigma(\lambda)}{\mathsf{d}\tau} = \\
&= \left(\frac{\mathsf{d}}{\mathsf{d}\tau}\left(\frac{\tau}{1+\tau}\right)\right)\left(\theta_x^{*\top}\Sigma_{xy}A(\tau)^{-1}\Sigma_{xy}^\top\theta_x^*\right) + \left(\frac{\tau}{1+\tau}\right)\left(\frac{\mathsf{d}}{\mathsf{d}\tau}\left(\theta_x^{*\top}\Sigma_{xy}A(\tau)^{-1}\Sigma_{xy}^\top\theta_x^*\right)\right) \\
&= \frac{1}{(1+\tau)^2}\,\theta_x^{*\top}\Sigma_{xy}A(\tau)^{-1}\Sigma_{xy}^\top\theta_x^* + \left(\frac{\tau}{1+\tau}\right)\left(\theta_x^{*\top}\Sigma_{xy}A(\tau)^{-1}\left(-\frac{\mathsf{d}}{\mathsf{d}\tau}A(\tau)\right)A(\tau)^{-1}\Sigma_{xy}^\top\theta_x^*\right) \\
&= \frac{1}{(1+\tau)^2}\theta_x^{*\top}\Sigma_{xy}A(\tau)^{-1}\Sigma_{xy}^\top\theta_x^* - \frac{\tau}{1+\tau}\left(\theta_x^{*\top}\Sigma_{xy}A(\tau)^{-1}\left(I + \frac{\Sigma_{xy}^\top\Sigma_{xy}}{(1+\tau)^2}\right)A(\tau)^{-1}\Sigma_{xy}^\top\theta_x^*\right) \\
&= \frac{1}{(1+\tau)}\theta_x^{*\top}\Sigma_{xy}A(\tau)^{-1}\left(\frac{A(\tau)}{1+\tau} - \tau\left(I + \frac{\Sigma_{xy}^\top\Sigma_{xy}}{(1+\tau)^2}\right)\right)A(\tau)^{-1}\Sigma_{xy}^\top\theta_x^* \\
&= \frac{1}{(1+\tau)}\theta_x^{*\top}\Sigma_{xy}A(\tau)^{-1}\left(\frac{\Sigma_{yy}+\tau I}{1+\tau} - \frac{\Sigma_{xy}^\top\Sigma_{xy}}{(1+\tau)^2} - \tau I - \tau\frac{\Sigma_{xy}^\top\Sigma_{xy}}{(1+\tau)^2}\right)A(\tau)^{-1}\Sigma_{xy}^\top\theta_x^* \\
&= \frac{1}{(1+\tau)^2}\theta_x^{*\top}\Sigma_{xy}A(\tau)^{-1}\left(\Sigma_{yy} - \Sigma_{xy}^\top\Sigma_{xy} - \tau^2 I\right)A(\tau)^{-1}\Sigma_{xy}^\top\theta_x^* \\
&= \frac{1}{(1+\tau)^2}\theta_x^{*\top}\Sigma_{xy}A(\tau)^{-1}\left(S_x^\Sigma - \tau^2 I\right)A(\tau)^{-1}\Sigma_{xy}^\top\theta_x^*,
\end{aligned}
\tag{G.35}
$$

where in the third line we used the identity $\frac{\mathsf{d}}{\mathsf{d}\tau}\left(A(\tau)^{-1}\right) = A(\tau)^{-1}\left(-\frac{\mathsf{d}}{\mathsf{d}\tau}A(\tau)\right)A(\tau)^{-1}$. Then, if $\tau \leq \sqrt{\lambda_{\min}\left(S_x^\Sigma\right)} = \tau_{\mathcal{C}}$, we have that $\left(S_x^\Sigma - \tau^2 I\right)$ is p.s.d., which in turn implies $\frac{\mathsf{d}C^\Sigma(\lambda)}{\mathsf{d}\tau} \geq 0$, thus giving the desired claim. The non-negativity of $\mathcal{C}^\Sigma(\lambda)$ readily follows from (G.34).

For the last statement, setting $\tau_{\mathcal{L}} = \lambda_{\min}(\Sigma)$ we show that $\mathcal{L}^\Sigma(\lambda)$ is monotonically increasing for all $\tau \in [\tau_{\mathcal{L}}, +\infty)$ as long as the additional bound on $2d/n$ holds. As $\lambda_{\min}\left(S_x^\Sigma\right) \leq \lambda_{\min}(\Sigma_{yy}) \leq \operatorname{tr}(\Sigma_{yy})/d = \operatorname{tr}(\Sigma - \Sigma_{xx})/d = 1$, we also have

$$
\tau_{\mathcal{C}} = \sqrt{\lambda_{\min}\left(S_x^\Sigma\right)} \geq \lambda_{\min}\left(S_x^\Sigma\right) \geq \lambda_{\min}(\Sigma) = \tau_{\mathcal{L}},
\tag{G.36}
$$

where the second inequality follows from Lemma G.3. Thus, from the monotonicity of $\tau(\lambda)$ in λ, the final result readily follows.

It remains to prove the monotonicity of $\mathcal{L}^\Sigma(\lambda)$ in $[\lambda_{\min}(\Sigma), +\infty)$. To do so, we again study the sign of (G.26).

For the first factor of the first term of (G.26), we have

$$
\begin{aligned}
\frac{\mathsf{d}}{\mathsf{d}\tau}\left(\tau^2\left\|(\Sigma+\tau I)^{-1}\Sigma^{1/2}\theta^*\right\|_2^2\right) &= 2\theta^{*\top}\left(\Sigma/\tau + I\right)^{-3}(\Sigma/\tau)^2\,\theta^* \\
&= 2\theta^{*\top}\Sigma^{1/2}(\Sigma+\tau I)^{-1}(\Sigma+\tau I)^{-1}\tau\Sigma(\Sigma+\tau I)^{-1}\Sigma^{1/2}\theta^* \\
&\geq \frac{2}{\tau}\lambda_{\min}\left(\Sigma(\Sigma+\tau)^{-1}\right)\tau^2\left\|(\Sigma+\tau I)^{-1}\Sigma^{1/2}\theta^*\right\|_2^2 \\
&= \frac{2}{\tau}\frac{\lambda_{\min}}{\lambda_{\min}+\tau}\tau^2\left\|(\Sigma+\tau I)^{-1}\Sigma^{1/2}\theta^*\right\|_2^2.
\end{aligned}
\tag{G.37}
$$

For the second factor of the first term of (G.26), we have

$$
1 - \frac{\operatorname{tr}\left((\Sigma+\tau I)^{-2}\Sigma^2\right)}{n} = 1 - \frac{1}{n}\sum_{k=1}^{2d}\frac{\lambda_k^2}{(\lambda_k+\tau)^2} \geq 1 - \frac{2d}{n}.
\tag{G.38}
$$

For the second factor of the second term of (G.26), we have

$$\begin{aligned}\frac{\mathsf{d}}{\mathsf{d}\tau}\left(1-\frac{\operatorname{tr}\left(\left(\Sigma+\tau I\right)^{-2}\Sigma^{2}\right)}{n}\right) &= \frac{2}{n}\sum_{k=1}^{2d}\frac{\lambda_k^2}{\left(\lambda_k+\tau\right)^3}\\ &\leq \frac{2}{n}\frac{1}{\tau\left(\lambda_{\min}+\tau\right)}\sum_{k=1}^{2d}\frac{\lambda_k^2}{\left(\lambda_k+\tau\right)}\\ &\leq \frac{4d}{n\tau\left(\lambda_{\min}+\tau\right)}.\end{aligned} \tag{G.39}$$

Thus, putting together (G.26), (G.37), (G.38), and (G.39), the monotonicity of $\mathcal{L}^{\Sigma}(\lambda)$ is implied by

$$\begin{aligned}&\frac{2}{\tau}\frac{\lambda_{\min}}{\lambda_{\min}+\tau}\tau^2\left\|\left(\Sigma+\tau I\right)^{-1}\Sigma^{1/2}\theta^*\right\|_2^2\left(1-\frac{2d}{n}\right)\overset{?}{\geq}\\ &\overset{?}{\geq}\left(\sigma^2+\tau^2\left\|\left(\Sigma+\tau I\right)^{-1}\Sigma^{1/2}\theta^*\right\|_2^2\right)\frac{4d}{n\tau\left(\lambda_{\min}+\tau\right)}\end{aligned} \tag{G.40}$$

Since we assumed that $2d/n \leq \lambda_{\min}/4 \leq 1/4$, we have

$$\frac{2}{\tau}\frac{\lambda_{\min}}{\lambda_{\min}+\tau}\left(1-\frac{2d}{n}\right)-\frac{4d}{n\tau\left(\lambda_{\min}+\tau\right)} = \frac{2}{\tau}\frac{\lambda_{\min}}{\lambda_{\min}+\tau}\left(1-\frac{2d}{n}-\frac{2d}{n\lambda_{\min}}\right)\geq\frac{1}{\tau}\frac{\lambda_{\min}}{\lambda_{\min}+\tau}, \tag{G.41}$$

and

$$\begin{aligned}\tau^2\left\|\left(\Sigma+\tau I\right)^{-1}\Sigma^{1/2}\theta^*\right\|_2^2 &\geq \lambda_{\min}\left(\tau^2\Sigma\left(\Sigma+\tau I\right)^{-2}\right)\\ &= \min_k\frac{\tau^2\lambda_k}{\left(\lambda_k+\tau\right)^2}\\ &= \min_k\frac{\lambda_k}{\left(\lambda_k/\tau+1\right)^2}\\ &\geq \min_k\frac{\lambda_k}{\left(\lambda_k/\lambda_{\min}+1\right)^2}\\ &= \lambda_{\min}\min_k\frac{\lambda_k/\lambda_{\min}}{\left(\lambda_k/\lambda_{\min}+1\right)^2}\\ &= \frac{\lambda_{\max}}{\left(\lambda_{\max}/\lambda_{\min}+1\right)^2},\end{aligned} \tag{G.42}$$

where in the fourth line we used that $\tau \geq \lambda_{\min}$, and in the last step we used that $f(x) := x/(x+1)^2$ is decreasing for $x \geq 1$.

Thus, using (G.41) and (G.42) gives that (G.40) is implied by

$$\frac{\lambda_{\max}}{\left(\lambda_{\max}/\lambda_{\min}+1\right)^2}\frac{1}{\tau}\frac{\lambda_{\min}}{\lambda_{\min}+\tau}\overset{?}{\geq}\sigma^2\frac{4d}{n\tau\left(\lambda_{\min}+\tau\right)}, \tag{G.43}$$

which holds since we assumed

$$\frac{2d}{n}\leq\frac{1}{2\sigma^2}\frac{\lambda_{\max}\lambda_{\min}}{\left(\lambda_{\max}/\lambda_{\min}+1\right)^2}. \tag{G.44}$$

□

G.1.1 Proofs on S_x^Σ

For completeness, in this section we prove two known results about S_x^Σ.

Lemma G.2. *Let $z = \left[x^\top, y^\top\right]^\top \sim \mathcal{P}_{XY}$ be distributed according to a mean-0, multivariate Gaussian distribution with covariance Σ, such that Σ is invertible. Then, the Schur complement S_x^Σ of Σ with respect to the top left block Σ_{xx} (see (6.14)) corresponds to the conditional covariance of y given x,* i.e.*,*

$$S_x^\Sigma = \mathrm{Cov}\left(y|x=\bar{x}\right) = \mathbb{E}_{y|x=\bar{x}}\left[\left(y - \mathbb{E}_{y|x=\bar{x}}[y]\right)\left(y - \mathbb{E}_{y|x=\bar{x}}[y]\right)^\top\right]. \tag{G.45}$$

Proof. Consider the expression $z^\top\Sigma^{-1}z$. According to the notation in (6.14) and in (G.17), we have

$$z^\top\Sigma^{-1}z = x^\top\left[\Sigma^{-1}\right]_{xx}x + y^\top\left[\Sigma^{-1}\right]_{yy}y + x^\top\left[\Sigma^{-1}\right]_{xy}y + y^\top\left[\Sigma^{-1}\right]_{yx}x. \tag{G.46}$$

Then, the formulas for the inverse of a block matrix give

$$\begin{aligned}
&z^\top\Sigma^{-1}z \\
&= x^\top\left(\Sigma_{xx}^{-1} + \Sigma_{xx}^{-1}\Sigma_{xy}S_x^{\Sigma^{-1}}\Sigma_{yx}\Sigma_{xx}^{-1}\right)x \\
&\quad + y^\top S_x^{\Sigma^{-1}}y + x^\top\left(-\Sigma_{xx}^{-1}\Sigma_{xy}S_x^{\Sigma^{-1}}\right)y + y^\top\left(-S_x^{\Sigma^{-1}}\Sigma_{yx}\Sigma_{xx}^{-1}\right)x \\
&= x^\top\Sigma_{xx}^{-1}x + \left(y - \Sigma_{yx}\Sigma_{xx}^{-1}x\right)^\top S_x^{\Sigma^{-1}}\left(y - \Sigma_{yx}\Sigma_{xx}^{-1}x\right).
\end{aligned} \tag{G.47}$$

Then, denoting with $p(x,y)$ and $p(x)$ the probability density functions of $z = \left[x^\top, y^\top\right]^\top$ and x respectively, we get that the probability density function of y conditioned on x takes the form

$$\begin{aligned}
p(y|x) &= \frac{p(x,y)}{p(x)} \\
&= \frac{\sqrt{(2\pi)^d\det\left(\Sigma_{xx}\right)}}{\sqrt{(2\pi)^{2d}\det\left(\Sigma\right)}}\frac{\exp\left(-\left[x^\top, y^\top\right]\Sigma^{-1}\left[x^\top, y^\top\right]^\top/2\right)}{\exp\left(-x^\top\Sigma_{xx}^{-1}x/2\right)} \\
&= \frac{1}{\sqrt{(2\pi)^d\det\left(S_x^\Sigma\right)}}\exp\left(-\left[x^\top, y^\top\right]\Sigma^{-1}\left[x^\top, y^\top\right]^\top/2 + x^\top\Sigma_{xx}^{-1}x/2\right) \\
&= \frac{\exp\left(-\left(y - \Sigma_{yx}\Sigma_{xx}^{-1}x\right)^\top S_x^{\Sigma^{-1}}\left(y - \Sigma_{yx}\Sigma_{xx}^{-1}x\right)/2\right)}{\sqrt{(2\pi)^d\det\left(S_x^\Sigma\right)}},
\end{aligned} \tag{G.48}$$

where we used Schur formula for the determinants in the third line, and (G.47) in the last step. Thus, we have that $p(y|x)$ describes the density of a multivariate Gaussian random variable, with covariance S_x^Σ. □

Lemma G.3. *Let $\Sigma \in \mathbb{R}^{2d\times 2d}$ be a p.s.d., invertible matrix. Then, the Schur complement $S_x^\Sigma \in \mathbb{R}^{d\times d}$ of Σ with respect to the top left block Σ_{xx} (see (6.14)) is such that*

$$\lambda_{\min}\left(S_x^\Sigma\right) \geq \lambda_{\min}\left(\Sigma\right). \tag{G.49}$$

Proof. Let $\Gamma \in \mathbb{R}^{2d\times d}$ be the rank-d matrix defined as

$$\Gamma = \begin{pmatrix} \Sigma_{xx}^{1/2} \\ \Sigma_{yx}\Sigma_{xx}^{-1/2} \end{pmatrix}, \tag{G.50}$$

and $S \in \mathbb{R}^{2d\times 2d}$ as the matrix containing S_x^Σ in its bottom-right $d \times d$ block, and 0 everywhere else. Then, we have that

$$\Sigma = S + \Gamma\Gamma^\top, \tag{G.51}$$

where both S and $\Gamma\Gamma^\top$ are rank-d p.s.d. matrices.

Denoting by $\lambda_k(S)$ the k-th largest eigenvalue of S, by the Courant–Fischer–Weyl min-max principle, we can write

$$\lambda_k(S) = \max_{W,\,\dim(W)=k} \min_{u\in W,\,\|u\|_2=1} \left(u^\top S u\right), \tag{G.52}$$

where with W we denote a generic k-dimensional subspace of $\mathbb{R}^{2d}$. Thus, the desired result follows from

$$\begin{aligned}
\lambda_{\min}(\Sigma) &= \lambda_{\min}\left(S + \Gamma\Gamma^\top\right) \\
&= \min_{\|u\|_2=1} u^\top \left(S + \Gamma\Gamma^\top\right) u \\
&\le \min_{u\in\ker\left(\Gamma\Gamma^\top\right),\,\|u\|_2=1} u^\top \left(S + \Gamma\Gamma^\top\right) u \\
&= \min_{u\in\ker\left(\Gamma\Gamma^\top\right),\,\|u\|_2=1} u^\top S u \\
&\le \max_{W,\,\dim(W)=d} \min_{u\in W,\,\|u\|_2=1} u^\top S u \\
&= \lambda_d(S) \\
&= \lambda_{\min}\left(S_x^\Sigma\right),
\end{aligned} \tag{G.53}$$

where the last step holds since the d smallest eigenvalues of S are equal to 0, and the d largest correspond to the ones of S_x^Σ. □

G.1.2 Remarks on Assumption 6.1

Our results on linear regression rely on Assumption 6.1, and in particular on the training samples to be normally distributed. This assumption is made for technical convenience, as the concentration results in Theorem 6.3 and Proposition 6.5 still hold under the following milder requirement.

Assumption G.4 (Data distribution)**.** *The input samples $\{z_i\}_{i=1}^n$ are n i.i.d. samples from a mean-0, sub-Gaussian distribution $\mathcal{P}_{XY}$, such that*

1. *its covariance $\Sigma \in \mathbb{R}^{2d\times 2d}$ is invertible, with $\lambda_{\max}(\Sigma) = O(1)$, $\lambda_{\min}(\Sigma) = \Omega(1)$, and $\operatorname{tr}(\Sigma) = 2d$;*
2. *for $z \sim P_Z$, the random variable $\Sigma^{-1/2}z$ has independent, mean-0, unit variance, sub-Gaussian entries.*

This assumption resembles the requirements A-B in Section 2.2 in Han and Xu [2023], where we also included the scaling of the trace. To formally state the equivalent of Theorem 6.3 and Proposition 6.5, one also has to enforce the following technical condition on the true parameter θ^*.

Assumption G.5. *Let $\delta = 1/72$, then we assume that*

$$\theta^* \text{ s.t. } \left\|\Sigma^{1/2}\tau(\Sigma+\tau I)^{-1}\theta^*\right\|_\infty \leq Cd^{\delta-1/2}. \tag{G.54}$$

In Proposition 10.3 in Han and Xu [2023], it is shown that this condition excludes a negligible fraction ($Ce^{-n^{2\delta}/C}$) of the θ^* on the unit ball. Since we set $\delta = 1/72$, following the same arguments of the proofs of Theorem 6.3 and Proposition 6.5, we have that Theorems 2.4 and 3.1 in Han and Xu [2023] imply the results below.

Theorem G.6. *Let Assumptions 6.9 and G.5 hold, and let $n = \Theta(d)$. Let $\hat{\theta}_{\mathsf{LR}}(\lambda)$ be defined as in* (6.8)*, and let $\mathcal{C}(\hat{\theta}_{\mathsf{LR}}(\lambda))$ be the amount of spurious correlations learned by the model $f_{\mathsf{LR}}\left(\hat{\theta}_{\mathsf{LR}}(\lambda),\cdot\right)$ as defined in* (6.3)*. Then, for any $\lambda > 0$, we have that, for every $t \in (0, 1/2)$,*

$$\mathbb{P}_{Z,G}\left(\left|\mathcal{C}(\hat{\theta}_{\mathsf{LR}}(\lambda)) - \mathcal{C}^{\Sigma}(\lambda)\right| \geq t\right) \leq Ct^{-13}d^{-1/8}, \tag{G.55}$$

where $\mathcal{C}^{\Sigma}(\lambda)$ is defined in (6.12)*, and C is a an absolute constant.*

Proposition G.7. *Let Assumptions 6.9 and G.5 hold, and let $n = \Theta(d)$. Let $\hat{\theta}_{\mathsf{LR}}(\lambda)$ be defined as in* (6.8)*, and let $\mathcal{L}(\hat{\theta}_{\mathsf{LR}}(\lambda))$ be the in-distribution test loss of the model $f_{\mathsf{LR}}(\hat{\theta}_{\mathsf{LR}}(\lambda),\cdot)$ as defined in* (6.5)*. Then, , for any $\lambda > 0$, we have that, for every $t \in (0, 1/2)$,*

$$\mathbb{P}_{Z,G}\left(\left|\mathcal{L}(\hat{\theta}_{\mathsf{LR}}(\lambda)) - \mathcal{L}^{\Sigma}(\lambda)\right| \geq t\right) \leq Ct^{-c}d^{-1/6.5}, \tag{G.56}$$

where $\mathcal{L}^{\Sigma}(\lambda)$ is defined in (6.20)*, and C and c are positive absolute constants.*

G.2 Proofs for Random Features

Lemma G.8. *We have that*

$$\|V\|_{\mathsf{op}} = O\left(\sqrt{\frac{p}{d}}\right), \tag{G.57}$$

$$\|Z\|_{\mathsf{op}} = O\left(\sqrt{d}\right), \tag{G.58}$$

with probability at least $1 - 2\exp(-cd)$ over V and Z, where c is an absolute constant. Furthermore, for every $i \in [n]$, we have

$$\left\|\|z_i\|_2 - \sqrt{2d}\right\|_{\psi_2} = O(1). \tag{G.59}$$

Proof. V has independent, mean-0, unit variance, sub-Gaussian entries. Then, the first statement is a direct consequences of Theorem 4.4.5 of Vershynin [2018] and of the scaling $d = o(p)$.

By Assumption 6.9, we have that Z has i.i.d. mean-0, Lipschitz concentrated rows. This property also implies that the rows are i.i.d. sub-Gaussian. Thus, by Remark 5.40 in Vershynin [2012], we have that

$$\left\|Z^\top Z - n\Sigma\right\|_{\mathsf{op}} = O\left(n\frac{d}{n}\right) = O(d), \tag{G.60}$$

with probability at least $1 - 2\exp(-c_1 d)$. Then, conditioning on this high probability event, by Weyl's inequality, we have

$$\left\|Z^\top Z\right\|_{\mathsf{op}} \leq \|n\Sigma\|_{\mathsf{op}} + \left\|Z^\top Z - n\Sigma\right\|_{\mathsf{op}} = O(d), \tag{G.61}$$

where the last step follows from the argument used to prove $\|\Sigma\|_{\mathsf{op}} = O(1)$ in Lemma C.1 in Bombari and Mondelli [2025a].

For the last statement, we have

$$2d = \operatorname{tr}(\Sigma) = \operatorname{tr}\left(\mathbb{E}\left[zz^\top\right]\right) = \mathbb{E}\left[\operatorname{tr}\left(zz^\top\right)\right] = \mathbb{E}\left[\operatorname{tr}\left(z^\top z\right)\right] = \mathbb{E}\left[\|z\|_2^2\right], \tag{G.62}$$

where we used the cyclic property of the trace. Furthermore, we have

$$\left\|\|z\|_2 - \mathbb{E}\left[\|z\|_2\right]\right\|_{\psi_2} = O(1), \tag{G.63}$$

since z is Lipschitz concentrated. Then,

$$0 \le 2d - \mathbb{E}\left[\|z\|_2\right]^2 = \mathbb{E}\left[\left(\|z\|_2 - \mathbb{E}\left[\|z\|_2\right]\right)^2\right] \le C_1, \tag{G.64}$$

for some absolute constant C_1. Thus, as $\sqrt{1-x} \ge 1 - x$ for $x \in [0,1]$, we obtain

$$1 - \frac{C_1}{2d} \le \sqrt{1 - \frac{C_1}{2d}} \le \frac{\mathbb{E}\left[\|z\|_2\right]}{\sqrt{2d}} \le 1. \tag{G.65}$$

Plugging this last result in (G.63) gives the desired claim. □

Lemma G.9. *We have that, denoting with $\tilde{\mu}^2 = \sum_{k\ge 2}\mu_k^2$, with μ_k denoting the k-th Hermite coefficient of ϕ,*

$$\left\|\mathbb{E}_V\left[\Phi\Phi^\top\right] - p\left(\mu_1^2\frac{ZZ^\top}{2d} + \tilde{\mu}^2 I\right)\right\|_{\mathsf{op}} = O\left(\frac{p\log^3 d}{\sqrt{d}}\right), \tag{G.66}$$

with probability at least $1 - 2\exp\left(-c\log^2 d\right)$ over Z, where c is an absolute constant.

Proof. For all $i \in [n]$, we define the functions $\phi^{(i)} : \mathbb{R} \to \mathbb{R}$ as $\phi^{(i)}(\cdot) = \phi(\|z_i\|_2 \cdot/\sqrt{2d})$. Note that $\phi^{(i)}$ is odd, since ϕ is odd by Assumption 6.7. Thus, denoting with $\mu_k^{(i)}$ the k-th Hermite coefficient of $\phi^{(i)}$, for every $i \in [n]$, we have that $\mu_k^{(i)} = 0$ for all even k. This implies that, by denoting with v a random vector distributed as the rows of V, *i.e.*, $\sqrt{2d}\,v$ is a standard Gaussian vector, we have

$$\begin{aligned}
\left[\mathbb{E}_V\left[\Phi\Phi^\top\right]\right]_{ij} &= p\,\mathbb{E}_v\left[\phi(z_i^\top v)\phi(z_j^\top v)\right] \\
&= p\,\mathbb{E}_v\left[\phi^{(i)}\left(\frac{z_i^\top}{\|z_i\|_2}\sqrt{2d}v\right)\phi^{(j)}\left(\frac{z_j^\top}{\|z_j\|_2}\sqrt{2d}v\right)\right] \\
&= p\sum_{k=0}^{+\infty}\mu_k^{(i)}\mu_k^{(j)}\left(\frac{z_i^\top z_j}{\|z_i\|_2\|z_j\|_2}\right)^k \\
&= p\mu_1^{(i)}\mu_1^{(j)}\frac{z_i^\top z_j}{\|z_i\|_2\|z_j\|_2} + p\sum_{k\ge 3}\mu_k^{(i)}\mu_k^{(j)}\left(\frac{z_i^\top z_j}{\|z_i\|_2\|z_j\|_2}\right)^k.
\end{aligned} \tag{G.67}$$

Then, denoting with $D_k \in \mathbb{R}^{n\times n}$ the diagonal matrix containing $\mu_k^{(i)}/\|z_i\|_2^k$ in its i-th entry, we can write

$$\mathbb{E}_V\left[\Phi\Phi^\top\right] = pD_1ZZ^\top D_1 + p\sum_{k\ge 3}D_k\left(ZZ^\top\right)^{\circ k}D_k. \tag{G.68}$$

Notice that, due to the last statement in Lemma G.8, we have that, jointly for all $i \in [n]$,

$$\left| \frac{\|z_i\|_2}{\sqrt{2d}} - 1 \right| = O\left(\frac{\log d}{\sqrt{d}} \right), \tag{G.69}$$

with probability at least $1 - 2\exp(-c_1 \log^2 d)$. Then, conditioning on such high probability event and denoting with ρ a standard Gaussian random variable, for all $i \in [n]$ we have

$$\begin{aligned}
\left| \mu_1^{(i)} - \mu_1 \right| &= \left| \mathbb{E}_\rho \left[\rho \phi^{(i)}(\rho) \right] - \mathbb{E}_\rho \left[\rho \phi(\rho) \right] \right| \\
&= \left| \mathbb{E}_\rho \left[\rho \left(\phi\left(\frac{\|z_i\|_2}{\sqrt{2d}} \rho \right) - \phi(\rho) \right) \right] \right| \\
&= \left| \mathbb{E}_\rho \left[\rho \left(\phi\left(\left(\frac{\|z_i\|_2}{\sqrt{2d}} - 1 \right) \rho + \rho \right) - \phi(\rho) \right) \right] \right| \\
&\leq \mathbb{E}_\rho \left[|\rho| \left| \phi\left(\left(\frac{\|z_i\|_2}{\sqrt{2d}} - 1 \right) \rho + \rho \right) - \phi(\rho) \right| \right] \\
&\leq L \mathbb{E}_\rho \left[|\rho| \left| \frac{\|z_i\|_2}{\sqrt{2d}} - 1 \right| |\rho| \right] \\
&= L \left| \frac{\|z_i\|_2}{\sqrt{2d}} - 1 \right| \mathbb{E}_\rho \left[\rho^2 \right] \\
&= O\left(\frac{\log d}{\sqrt{d}} \right),
\end{aligned} \tag{G.70}$$

where we used Jensen's inequality in the fourth line, the L-Lipschitzness of ϕ in the fifth line, and (G.69) in the last step. With a similar approach, denoting with $\|\cdot\|_{L^2}$ the L^2 norm with respect to the Gaussian measure, we have that for all $i \in [n]$

$$\begin{aligned}
\left| \left\| \phi^{(i)} \right\|_{L^2} - \|\phi\|_{L^2} \right| &\leq \left\| \phi^{(i)} - \phi \right\|_{L^2} \\
&= \mathbb{E}_\rho \left[\left(\phi^{(i)}(\rho) - \phi(\rho) \right)^2 \right]^{1/2} \\
&= \mathbb{E}_\rho \left[\left(\phi\left(\left(\frac{\|z_i\|_2}{\sqrt{2d}} - 1 \right) \rho + \rho \right) - \phi(\rho) \right)^2 \right]^{1/2} \\
&\leq L \left| \frac{\|z_i\|_2}{\sqrt{2d}} - 1 \right| \mathbb{E}_\rho \left[\rho^2 \right]^{1/2} \\
&= O\left(\frac{\log d}{\sqrt{d}} \right),
\end{aligned} \tag{G.71}$$

which directly implies that, for all $i \in [n]$, $\left\| \phi^{(i)} \right\|_{L^2} = \sum_{k \geq 0} \left(\mu_k^{(i)} \right)^2 = \Theta(1)$, and that

$$\left| \sum_{k \geq 3} \left(\mu_k^{(i)} \right)^2 - \sum_{k \geq 3} \mu_k^2 \right| \leq \left| \left\| \phi^{(i)} \right\|_{L^2}^2 - \|\phi\|_{L^2}^2 \right| + \left| \left(\mu_1^{(i)} \right)^2 - \mu_1^2 \right| = O\left(\frac{\log d}{\sqrt{d}} \right). \tag{G.72}$$

Thus, we are ready to estimate the operator norm of the off-diagonal part of the second term

on the RHS of (G.68), specifically

$$\begin{aligned}
&\left\|\sum_{k\geq 3} D_k \left(ZZ^\top\right)^{\circ k} D_k - \operatorname{diag}\left(\sum_{k\geq 3} D_k \left(ZZ^\top\right)^{\circ k} D_k\right)\right\|_{\mathsf{op}} \\
&\quad \leq \sum_{k\geq 3} \left\| D_k \left(ZZ^\top\right)^{\circ k} D_k - \operatorname{diag}\left(D_k \left(ZZ^\top\right)^{\circ k} D_k\right)\right\|_F \\
&\quad \leq \sum_{k\geq 3} \max_{i\neq j} \left(\frac{\left|z_i^\top z_j\right|}{\|z_i\|_2 \|z_j\|_2}\right)^k \left(\sum_{i\in[n], j\in[n]} \left(\mu_k^{(i)} \mu_k^{(j)}\right)^2\right)^{1/2} \\
&\quad \leq \max_{i\neq j} \left(\frac{\left|z_i^\top z_j\right|}{\|z_i\|_2 \|z_j\|_2}\right)^3 \sum_{i=0}^{n} \sum_{k\geq 3} \left(\mu_k^{(i)}\right)^2 \\
&\quad = O\left(\frac{1}{d^{3/2}} \log^3 d\, n\right) = O\left(\frac{\log^3 d}{\sqrt{d}}\right),
\end{aligned} \tag{G.73}$$

where in the first step we replaced the operator norm with the Frobenius norm, and used triangle inequality; in the fifth line we used that $\|z_i\|_2 = \Theta(\sqrt{d})$ for all $i \in [n]$ (true because of (G.69)), and that jointly for all $i \neq j$ we have $\left|z_i^\top z_j\right| / \|z_j\|_2 = O(\log d)$ with probability at least $1 - 2\exp(-c_2 \log^2 d)$ since the z_i-s are independent sub-Gaussian vectors (since they are mean-0 and Lipschitz concentrated). The diagonal part of the second term on the RHS of (G.68) respects

$$\left\|\operatorname{diag}\left(\sum_{k\geq 3} D_k \left(ZZ^\top\right)^{\circ k} D_k\right) - \tilde{\mu}^2 I\right\|_{\mathsf{op}} = \max_{i\in[n]} \left|\sum_{k\geq 3} \left(\mu_k^{(i)}\right)^2 - \sum_{k\geq 3} \mu_k^2\right| = O\left(\frac{\log d}{\sqrt{d}}\right), \tag{G.74}$$

because of (G.72). Lastly, notice that,

$$\begin{aligned}
&\left\|D_1 Z Z^\top D_1 - \mu_1^2 \frac{ZZ^\top}{2d}\right\|_{\mathsf{op}} = \\
&= \sup_{\|u\|_2=1} \left|u^\top D_1 Z Z^\top D_1 u - \mu_1^2 u^\top \frac{ZZ^\top}{2d} u\right| \\
&= \sup_{\|u\|_2=1} \left|\left\|Z^\top D_1 u\right\|_2^2 - \mu_1^2 \left\|\frac{Z^\top}{\sqrt{2d}} u\right\|_2^2\right| \\
&\leq \sup_{\|u\|_2=1} \left(\left\|Z^\top D_1 u\right\|_2 + \mu_1 \left\|\frac{Z^\top}{\sqrt{2d}} u\right\|_2\right) \sup_{\|u\|_2=1} \left(\left\|Z^\top D_1 u - \mu_1 \frac{Z^\top}{\sqrt{2d}} u\right\|_2\right) \\
&\leq \left(\left\|Z^\top D_1\right\|_{\mathsf{op}} + \mu_1 \left\|\frac{Z^\top}{\sqrt{2d}}\right\|_{\mathsf{op}}\right) \left\|Z^\top D_1 - \mu_1 \frac{Z^\top}{\sqrt{2d}}\right\|_{\mathsf{op}} \\
&\leq \left(\|Z\|_{\mathsf{op}} \|D_1\|_{\mathsf{op}} + \mu_1 \frac{\|Z\|_{\mathsf{op}}}{\sqrt{2d}}\right) \|Z\|_{\mathsf{op}} \left\|D_1 - \frac{\mu_1}{\sqrt{2d}}\right\|_{\mathsf{op}}.
\end{aligned} \tag{G.75}$$

By Lemma G.8, we have that $\|Z\|_{\mathsf{op}} = O\left(\sqrt{d}\right)$ with probability at least $1 - 2\exp(-c_2 d)$, and since $\|z_i\|_2 = \Theta(\sqrt{d})$ and $\mu_1^{(i)} = O(1)$ for all $i \in [n]$ (true because of (G.69) and (G.70)

respectively), we have that $\|D_1\|_{\mathsf{op}} = O\left(1/\sqrt{d}\right)$. Furthermore, we have

$$\begin{aligned}\left\|D_1 - \frac{\mu_1}{\sqrt{2d}}\right\|_{\mathsf{op}} &= \max_i \left|\frac{\mu_1^{(i)}}{\|z_i\|_2} - \frac{\mu_1}{\sqrt{2d}}\right| \\ &\leq \max_i \frac{1}{\|z_i\|_2}\left(\left|\mu_1^{(i)} - \mu_1\right| + \mu_1\left|1 - \frac{\|z_i\|_2}{\sqrt{2d}}\right|\right) \\ &= O\left(\frac{\log d}{d}\right),\end{aligned} \tag{G.76}$$

where the last step is a consequence of (G.69) and (G.70). Then, we have that (G.75) reads

$$\left\|D_1 Z Z^\top D_1 - \mu_1^2 \frac{Z Z^\top}{2d}\right\|_{\mathsf{op}} = O\left(\frac{\log d}{\sqrt{d}}\right). \tag{G.77}$$

A standard application of the triangle inequality to (G.73), (G.74) and (G.77) gives

$$\left\|\left(D_1 Z Z^\top D_1 + \sum_{k\geq 3} D_k \left(Z Z^\top\right)^{\circ k} D_k\right) - \left(\mu_1^2 \frac{Z Z^\top}{2d} + \tilde{\mu}^2 I\right)\right\|_{\mathsf{op}} = O\left(\frac{\log^3 d}{\sqrt{d}}\right), \tag{G.78}$$

with probability at least $1 - 2\exp\left(-c_3 \log^2 d\right)$ over Z (where we used $\mu_2 = 0$ since ϕ is odd), which readily gives the thesis when plugged in (G.68). □

Lemma G.10. *We have that*

$$\|\Phi\|_{\mathsf{op}} = O\left(\sqrt{p}\right), \tag{G.79}$$

$$\left\|\Phi\Phi^\top - \mathbb{E}_V\left[\Phi\Phi^\top\right]\right\|_{\mathsf{op}} = O\left(\sqrt{pd}\right), \tag{G.80}$$

$$\lambda_{\min}\left(\Phi\Phi^\top\right) = \Omega\left(p\right), \tag{G.81}$$

with probability at least $1 - 2\exp\left(-c\log^2 d\right)$ *over* Z *and* V*, where* c *is an absolute constant.*

Proof. $\Phi^\top$ is a matrix with i.i.d. rows in the probability space of V. In particular, its i-th row takes the form

$$\left[\Phi^\top\right]_{i:} = \phi\left(Z V_{i:}\right) = \phi\left(Z V_{i:}\right) - \mathbb{E}_V\left[\phi\left(Z V_{i:}\right)\right], \tag{G.82}$$

where the last step holds since the (Gaussian) distribution of $V_{i:}$ is symmetric and ϕ is an odd function by Assumption 6.7. Then, since $\sqrt{2d}\, V_{i:}$ is a standard Gaussian (and hence Lipschitz concentrated) random vector, and ϕ is a Lipschitz continuous function, we have that

$$\left\|\left[\Phi^\top\right]_{i:}\right\|_{\psi_2} = O\left(\frac{\|Z\|_{\mathsf{op}}}{\sqrt{d}}\right) = O\left(1\right), \tag{G.83}$$

where the $\|\cdot\|_{\psi_2}$ is meant on the probability space of V, and the second step holds with probability at least $1 - 2\exp\left(-c_1 d\right)$ over Z due to Lemma G.8. Conditioning on this high probability event, $\Phi^\top$ is a $p \times n$ matrix whose rows are i.i.d. mean-0 sub-Gaussian random vectors in $\mathbb{R}^n$. Then, by Lemma B.7 in Bombari et al. [2022b], we have

$$\left\|\Phi^\top\right\|_{\mathsf{op}} = O\left(\sqrt{n} + \sqrt{p}\right) = O\left(\sqrt{p}\right), \tag{G.84}$$

with probability at least $1 - 2\exp(-c_2 n)$ over V, where the second step holds because $n = o(p)$.

For the second part of the proof, we again follow the argument in Lemma B.7 in Bombari et al. [2022b], which in turn exploits the discussion in Remark 5.40 in Vershynin [2012], and conclude that

$$\left\|\Phi\Phi^\top - \mathbb{E}_V\left[\Phi\Phi^\top\right]\right\|_{\mathsf{op}} = O\left(p\sqrt{\frac{n}{p}}\right) = O\left(\sqrt{pn}\right) = O\left(\sqrt{pd}\right), \tag{G.85}$$

with probability at least $1 - 2\exp(-c_3 n)$ over Z and V.

For the last statement, Lemma G.9 and Weyl's inequality imply that, with probability at least $1 - 2\exp\left(-c_2 \log^2 d\right)$ over Z we have

$$\begin{aligned}\lambda_{\min}\left(\Phi\Phi^\top\right) &\geq p\,\lambda_{\min}\left(\mu_1^2\frac{ZZ^\top}{2d} + \tilde{\mu}^2 I\right) - \left\|\mathbb{E}_V\left[\Phi\Phi^\top\right] - p\left(\mu_1^2\frac{ZZ^\top}{2d} + \tilde{\mu}^2 I\right)\right\|_{\mathsf{op}} \\ &\quad - \left\|\Phi\Phi^\top - \mathbb{E}_V\left[\Phi\Phi^\top\right]\right\|_{\mathsf{op}} \\ &\geq p\tilde{\mu}^2 - O\left(\frac{p\log^3 d}{\sqrt{d}}\right) - O\left(\sqrt{pd}\right) = \Omega(p),\end{aligned} \tag{G.86}$$

where the last step is true since $\tilde{\mu} \neq 0$, as ϕ is non-linear by Assumption 6.7. □

Lemma G.11. *Let $\tilde{\phi} : \mathbb{R} \to \mathbb{R}$ be defined as $\tilde{\phi}(\cdot) := \phi(\cdot) - \mu_1(\cdot)$, and set*

$$n' = \min\left(\left\lfloor\frac{p}{\log^4 p}\right\rfloor, \left\lfloor\frac{d^{3/2}}{\log^3 d}\right\rfloor\right). \tag{G.87}$$

Let $\{\hat{z}_i\}_{i=1}^{n'}$ be n' i.i.d. random variables sampled from a distribution respecting Assumption 6.9, not necessarily with the same covariance as $\mathcal{P}_{XY}$, and independent from V. Then, if $\tilde{\Phi}_{n'} \in \mathbb{R}^{n'\times p}$ is defined as the matrix containing $\tilde{\phi}(V\hat{z}_i)$ in its i-th row, we have that

$$\left\|\tilde{\Phi}_{n'}\right\|_{\mathsf{op}} = O\left(\sqrt{p}\right), \tag{G.88}$$

with probability at least $1 - 2\exp\left(-c\log^2 d\right)$ over $\{\hat{z}_i\}_{i=1}^{n'}$ and V, where c is an absolute constant.

Proof. The proof follows the same strategy as Lemma C.8 in Bombari and Mondelli [2025a], with the only difference that they work under their Assumption 1.2, *i.e.* that the data is normalized as $\|\hat{z}_i\|_2 = \sqrt{2d}$ (in our notation). This difference, however, does not affect the result. We can in fact condition on the high probability event that all $\hat{z}_i$ are such that $\|\hat{z}_i\|_2 = \Theta(\sqrt{d})$, which holds with probability at least $1 - 2\exp(-cd)$ by Lemma G.8, and proceed in the same way (as their Equation (C.78) now holds) until their Equation (C.81), which requires their Lemma C.7, *i.e.* that

$$\left\|\mathbb{E}_V\left[\tilde{\Phi}_{n'}\tilde{\Phi}_{n'}{}^\top\right]\right\|_{\mathsf{op}} = O\left(p\right). \tag{G.89}$$

This holds also in our case, as it can be proven following the argument in (G.73) and (G.74), where now the n in the last line of (G.73) has to be replaced with n', making the RHS there being $O(1)$, as $n' = O\left(\frac{d^{3/2}}{\log^3 d}\right)$ by definition. Lastly, the normalization of the data is used one

more time in their Equation (C.91), but it is not critical to obtain the result, as $\|\hat{z}_i\|_2 = \Theta(\sqrt{d})$ is sufficient. We remark that Assumption 1.2 in Bombari and Mondelli [2025a] also requires the covariance of the distribution to be well-conditioned, which however is not required for the purposes of the above mentioned lemmas. □

Lemma G.12. *Let $\tilde{\phi} : \mathbb{R} \to \mathbb{R}$ be defined as $\tilde{\phi}(\cdot) := \phi(\cdot) - \mu_1(\cdot)$, and let $z \in \mathbb{R}^{2d}$ be sampled from a distribution respecting Assumption 6.9, not necessarily with the same covariance as $\mathcal{P}_{XY}$, and independent from V. Then we have that*

$$\left\|\mathbb{E}_z\left[\tilde{\phi}(Vz)\tilde{\phi}(Vz)^\top\right]\right\|_{\mathsf{op}} = O\left(\log^4 d + \frac{p\log^3 d}{d^{3/2}}\right), \tag{G.90}$$

with probability at least $1 - 2p^2\exp\left(-c\log^2 d\right)$ over V, where c is an absolute constant.

Proof. The proof follows a similar path as the one in Lemma C.15 in Bombari and Mondelli [2025a]. In particular, set

$$n' = \min\left(\left\lfloor\frac{p}{\log^4 p}\right\rfloor, \left\lfloor\frac{d^{3/2}}{\log^3 d}\right\rfloor\right), \qquad N = p^2 n', \tag{G.91}$$

and let $\tilde{\Phi}_N \in \mathbb{R}^{N\times p}$ be a matrix containing $\tilde{\phi}(V\hat{z}_i)$ in its i-th row, where every $\{\hat{z}_i\}_{i=1}^N$ is sampled independently from the same distribution of z. Thus, $\tilde{\Phi}_N$ can be seen as the vertical stacking of p^2 matrices with size $n' \times p$. All these matrices respect the hypotheses of Lemma G.11, and hence have their operator norm bounded by $O\left(\sqrt{p}\right)$ with probability at least $1 - 2\exp\left(-c_1\log^2 d\right)$. Thus, performing a union bound over these p^2 matrices, we get

$$\left\|\tilde{\Phi}_N^\top\tilde{\Phi}_N\right\|_{\mathsf{op}} = O\left(p^2\, p\right) = O\left(\frac{Np}{n'}\right) = O\left(N\log^4 p + \frac{Np\log^3 d}{d^{3/2}}\right), \tag{G.92}$$

with probability at least $1 - 2p^2\exp\left(-c_1\log^2 d\right)$ over V and $\{\hat{z}_i\}_{i=1}^N$.

Via the same argument used for the last statement of Lemma G.8, denoting with $v_k \in \mathbb{R}^{2d}$ the k-th row of V, we have that $\|v_k\|_2 = O(1)$ uniformly for every k with probability at least $1 - 2p\exp(-c_2 d)$. Conditioning on such event, we have that each entry of $\tilde{\phi}(V\hat{z}_1)$ is sub-Gaussian (with uniformly bounded sub-Gaussian norm), since $\hat{z}_1$ is sub-Gaussian (as it is mean-0 and Lipschitz concentrated) and $\tilde{\phi}$ is a Lipschitz function. Thus, we have that each entry of $\mathbb{E}_{\hat{z}_1}\left[\tilde{\phi}(V\hat{z}_1)\right]$ is $O(1)$ (see Proposition 2.5.2 in Vershynin [2018]), and therefore that $\left\|\mathbb{E}_{z_1}\left[\tilde{\phi}(V\hat{z}_1)\right]\right\|_{\psi_2} = O\left(\left\|\mathbb{E}_{z_1}\left[\tilde{\phi}(V\hat{z}_1)\right]\right\|_2\right) = O\left(\sqrt{p}\right)$. Then, conditioning on the high probability event $\|V\|_{\mathsf{op}} = O\left(\sqrt{p/d}\right)$ given by Lemma G.8, we have

$$\left\|\tilde{\phi}(V\hat{z}_1)\right\|_{\psi_2} \le \left\|\tilde{\phi}(V\hat{z}_1) - \mathbb{E}_{z_1}\left[\tilde{\phi}(V\hat{z}_1)\right]\right\|_{\psi_2} + \left\|\mathbb{E}_{z_1}\left[\tilde{\phi}(V\hat{z}_1)\right]\right\|_{\psi_2} = O\left(\sqrt{\frac{p}{d}} + \sqrt{p}\right) = O\left(\sqrt{p}\right), \tag{G.93}$$

where the second step holds because $\hat{z}_1$ is Lipschitz concentrated and $\tilde{\phi}$ is Lipschitz. Since the rows of $\tilde{\Phi}_N$ are identically distributed, this also holds jointly for all other $\hat{z}_i$-s, for $i \in [N]$. Then, $\tilde{\Phi}_N/\sqrt{p}$ is a matrix with independent sub-Gaussian rows, and by Theorem 5.39 in Vershynin [2012] (see their Remark 5.40 and Equation (5.25)), we have that

$$\frac{1}{p}\left\|\frac{\tilde{\Phi}_N^\top\tilde{\Phi}_N}{N} - \mathbb{E}_z\left[\tilde{\phi}(Vz)\tilde{\phi}(Vz)^\top\right]\right\|_{\mathsf{op}} = O\left(\sqrt{\frac{p}{N}}\right), \tag{G.94}$$

with probability at least $1 - 2\exp(-c_3 p)$ over $\{\hat{z}_i\}_{i=1}^N$. Then, we have

$$\begin{aligned}\left\|\mathbb{E}_z\left[\tilde{\phi}(Vz)\tilde{\phi}(Vz)^\top\right]\right\|_{\mathsf{op}} &\leq \left\|\frac{\tilde{\Phi}_N^\top\tilde{\Phi}_N}{N} - \mathbb{E}_z\left[\tilde{\phi}(Vz)\tilde{\phi}(Vz)^\top\right]\right\|_{\mathsf{op}} + \frac{\left\|\tilde{\Phi}_N^\top\tilde{\Phi}_N\right\|_{\mathsf{op}}}{N} \\ &= O\left(p\sqrt{\frac{p}{N}}\right) + O\left(\log^4 p + \frac{p\log^3 d}{d^{3/2}}\right) \\ &= O\left(\sqrt{p}\sqrt{\frac{\log^4 p}{p}} + \sqrt{p}\sqrt{\frac{\log^3 d}{d^{3/2}}}\right) + O\left(\log^4 p + \frac{p\log^3 d}{d^{3/2}}\right) \\ &= O\left(\log^4 p + \frac{p\log^3 d}{d^{3/2}}\right),\end{aligned} \tag{G.95}$$

where the first step follows from the triangle inequality, the second step is a consequence of (G.94) and (G.92), and the third step follows from the definition of N.

Taking the intersection between the high probability events in (G.92), (G.93) and (G.94), the previous equation then holds with probability at least $1 - 2p^2\exp\left(-c_4\log^2 d\right)$ over V and $\{\hat{z}_i\}_{i=1}^N$. Also note that its LHS does not depend on $\{\hat{z}_i\}_{i=1}^N$, which were introduced as auxiliary random variables. Thus, the high probability bound holds restricted to the probability space of V, and the desired result follows. □

Lemma G.13. *Let $\tilde{\phi}:\mathbb{R}\to\mathbb{R}$ be defined as $\tilde{\phi}(\cdot) := \phi(\cdot) - \mu_1(\cdot)$ and $\tilde{\Phi}\in\mathbb{R}^{n\times p}$ as the matrix containing $\tilde{\phi}(Vz_i)$ in its i-th row. Then, we have*

$$\left\|\tilde{\Phi}V\right\|_{\mathsf{op}} = O\left(\sqrt{p}\log d + \frac{p\log d}{d}\right), \tag{G.96}$$

with probability at least $1 - 2\exp\left(-c\log^2 d\right)$ over Z and V, where c is an absolute constant.

Proof. Note that $\tilde{\phi}$ is Lipschitz (since ϕ is Lipschitz by Assumption 6.7). During all the proof, we condition on the event $\|Z\|_{\mathsf{op}} = O\left(\sqrt{d}\right)$ and $\|z_i\|_2 = \Theta\left(\sqrt{d}\right)$ for all $i\in[n]$, which holds with probability at least $1 - 2\exp(-c_1 d)$ by Lemma G.8. During the proof we also use the shorthand $v\in\mathbb{R}^{2d}$ to denote a random vector such that $\sqrt{2d}\,v$ is a standard Gaussian vector, *i.e.*, it has the same distribution as the rows of V. This implies

$$\mathbb{E}_v\left[\left\|\tilde{\phi}\left(Zv\right)\right\|_2\right] = O\left(\sqrt{n}\right), \qquad \left\|\left\|\tilde{\phi}\left(Zv\right)\right\|_2 - \mathbb{E}_v\left[\left\|\tilde{\phi}\left(Zv\right)\right\|_2\right]\right\|_{\psi_2} = O\left(1\right), \tag{G.97}$$

and

$$\mathbb{E}_v\left[\|v\|_2\right] = O\left(1\right), \qquad \left\|\|v\|_2 - \mathbb{E}_v\left[\|v\|_2\right]\right\|_{\psi_2} = O\left(\frac{1}{\sqrt{d}}\right), \tag{G.98}$$

where both sub-Gaussian norms are meant on the probability space of v, and where the very first equation follows from the discussion in Lemma C.3 in Bombari et al. [2022b]. Then, there exists an absolute constant C_1 such that we jointly have

$$\left\|\tilde{\phi}\left(Zv\right)\right\|_2 \leq C_1\sqrt{d}, \qquad \|v\|_2 \leq C_1 \tag{G.99}$$

with probability at least $1 - 2\exp(-c_2 d)$ over v.

Let E_k be the indicator defined on the high probability event above with respect to the random variable $v_k := V_{:k}$ (the k-th row of V), *i.e.*

$$E_k := \mathbf{1}\left(\|v_k\|_2 \leq C_1 \text{ and } \left\|\tilde{\phi}\left(Zv_k\right)\right\|_2 \leq C_1\sqrt{d}\right), \tag{G.100}$$

and we define $E \in \mathbb{R}^{p\times p}$ as the diagonal matrix containing E_k in its k-th entry. Notice that we have $\|I-E\|_{\mathsf{op}} = 0$ with probability at least $1 - 2p\exp\left(-c_2 d\right)$, and $\mathbb{E}_V\left[\|I-E\|_{\mathsf{op}}\right] \leq 2p\exp\left(-c_2 d\right)$.

Thus, we have

$$\begin{aligned}
\left\|\mathbb{E}_V\left[\tilde{\Phi}\left(I-E\right)V\right]\right\|_{\mathsf{op}} &\leq \mathbb{E}_V\left[\left\|\tilde{\Phi}\right\|_{\mathsf{op}}\|1-E\|_{\mathsf{op}}\|V\|_{\mathsf{op}}\right] \\
&\leq \mathbb{E}_V\left[\left\|\tilde{\Phi}\right\|_{\mathsf{op}}^2\|V\|_{\mathsf{op}}^2\right]^{1/2}\mathbb{E}_V\left[\|I-E\|_{\mathsf{op}}^2\right]^{1/2} \\
&\leq \mathbb{E}_V\left[\left\|\tilde{\Phi}\right\|_{\mathsf{op}}^4\right]^{1/4}\mathbb{E}_V\left[\|V\|_{\mathsf{op}}^4\right]^{1/4}\left(2p\exp\left(-c_2 d\right)\right)^{1/2} \\
&\leq \mathbb{E}_V\left[\left\|\tilde{\Phi}\right\|_{F}^4\right]^{1/4}\mathbb{E}_V\left[\|V\|_{F}^4\right]^{1/4}\left(2p\exp\left(-c_2 d\right)\right)^{1/2} \\
&= o(1),
\end{aligned} \tag{G.101}$$

where the last step holds because of our initial conditioning on Z: the first two terms are the sum of finite powers of sub-Gaussian random variables (the entries of $\tilde{\Phi}$ and V), and thus (see Proposition 2.5.2 in Vershynin [2018]) the first two factors in the third line of the previous equation will be $O\left(p^\alpha\right)$ for some finite α, which gives the last line due to Assumption 6.8.

As in Lemma G.9, we introduce the notation (for all $i \in [n]$) $\tilde{\phi}^{(i)} : \mathbb{R} \to \mathbb{R}$ such that $\tilde{\phi}^{(i)}(\cdot) = \tilde{\phi}(\|z_i\|_2 \cdot/\sqrt{2d})$. Thus, denoting with $v \in \mathbb{R}^{2d}$ a random vector such that $\sqrt{2d}v$ is standard Gaussian (*i.e.*, distributed as the rows of V), we can write

$$\begin{aligned}
\left[\mathbb{E}_V\left[\tilde{\Phi}V\right]\right]_{ij} &= p\left[\mathbb{E}_v\left[\tilde{\phi}(Zv)v^\top\right]\right]_{ij} \\
&= \frac{p}{\sqrt{2d}}\mathbb{E}_v\left[\tilde{\phi}^{(i)}\left(\frac{z_i^\top}{\|z_i\|_2}\sqrt{2d}v\right)\left(e_j^\top\left(\sqrt{2d}v\right)\right)\right] \\
&= \frac{p}{\sqrt{2d}}\frac{\tilde{\mu}_1^{(i)}z_i^\top e_j}{\|z_i\|_2},
\end{aligned} \tag{G.102}$$

where $\tilde{\mu}_1^{(i)}$ is the first Hermite coefficient of $\tilde{\phi}^{(i)}$. Then, denoting with $\tilde{D} \in \mathbb{R}^{n\times n}$ the diagonal matrix containing $\tilde{\mu}_1^{(i)}/\|z_i\|_2$ in its i-th entry, we can write

$$\left\|\mathbb{E}_V\left[\tilde{\Phi}V\right]\right\|_{\mathsf{op}} = \frac{p}{\sqrt{2d}}\left\|\tilde{D}Z\right\|_{\mathsf{op}} \leq \frac{p}{\sqrt{2d}}\left\|\tilde{D}\right\|_{\mathsf{op}}\|Z\|_{\mathsf{op}}. \tag{G.103}$$

Then, since we conditioned on $\|z_i\|_2 = \Theta\left(\sqrt{d}\right)$ for all $i \in [n]$, following the same argument as in (G.70), and since the first Hermite coefficient of $\tilde{\phi}$ is 0 by definition, we have

$$\left\|\mathbb{E}_V\left[\tilde{\Phi}V\right]\right\|_{\mathsf{op}} = O\left(\frac{p}{\sqrt{d}}\frac{\log d}{d}\sqrt{d}\right) = O\left(\frac{p\log d}{d}\right), \tag{G.104}$$

with probability at least $1 - 2\exp\left(-c_3\log^2 d\right)$ over Z. A standard application of the triangle inequality to this last equation and (G.101) then gives

$$\left\|\mathbb{E}_V\left[\tilde{\Phi}EV\right]\right\|_{\mathsf{op}} \leq \left\|\mathbb{E}_V\left[\tilde{\Phi}V\right]\right\|_{\mathsf{op}} + \left\|\mathbb{E}_V\left[\tilde{\Phi}\left(I-E\right)V\right]\right\|_{\mathsf{op}} = O\left(\frac{p\log d}{d}\right), \tag{G.105}$$

with probability at least $1 - 2\exp\left(-c_4 \log^2 d\right)$ over Z.

Let's now look at

$$\tilde{\Phi}EV - \mathbb{E}_V\left[\tilde{\Phi}EV\right] = \sum_{k=1}^{p} \tilde{\phi}(Zv_k)E_k v_k^\top - \mathbb{E}_{v_k}\left[\tilde{\phi}(Zv_k)E_k v_k^\top\right] =: \sum_{k=1}^{p} W_k, \tag{G.106}$$

where we defined the shorthand $W_k = \tilde{\phi}(Zv_k)E_k v_k^\top - \mathbb{E}_{v_k}\left[\tilde{\phi}(Zv_k)E_k v_k^\top\right]$. (G.106) is the sum of p i.i.d. mean-0 random matrices W_k (in the probability space of V), such that

$$\begin{aligned}\sup_{v_k}\left\|\tilde{\phi}(Zv_k)E_k v_k^\top - \mathbb{E}_{v_k}\left[\tilde{\phi}(Zv_k)E_k v_k^\top\right]\right\|_{\mathsf{op}} &\leq 2\sup_{v_k}\left\|\tilde{\phi}(Zv_k)E_k v_k^\top\right\|_{\mathsf{op}} \\ &= 2\sup_{v_k}\left\|\tilde{\phi}(Zv_k)\right\|_2 \|v_k\|_2 E_k \\ &\leq 2C_1^2\sqrt{d},\end{aligned} \tag{G.107}$$

because of (G.100). Then, by matrix Bernstein's inequality for rectangular matrices (see Exercise 5.4.15 in Vershynin [2018]), we have that

$$\mathbb{P}_V\left(\left\|\tilde{\Phi}EV - \mathbb{E}_V\left[\tilde{\Phi}EV\right]\right\|_{\mathsf{op}} \geq t\right) \leq (n+d)\exp\left(-\frac{t^2/2}{\sigma^2 + 2C_1^2\sqrt{d}t/3}\right), \tag{G.108}$$

where σ^2 is defined as

$$\sigma^2 = p\max\left(\left\|\mathbb{E}_{v_k}\left[W_k W_k^\top\right]\right\|_{\mathsf{op}}, \left\|\mathbb{E}_{v_k}\left[W_k^\top W_k\right]\right\|_{\mathsf{op}}\right). \tag{G.109}$$

For every matrix A, we have $\mathbb{E}\left[(A - \mathbb{E}[A])(A - \mathbb{E}[A])^\top\right] = \mathbb{E}\left[AA^\top\right] - \mathbb{E}[A]\mathbb{E}[A]^\top \preceq \mathbb{E}\left[AA^\top\right]$. Thus,

$$\begin{aligned}\left\|\mathbb{E}_{v_k}\left[W_k W_k^\top\right]\right\|_{\mathsf{op}} &\leq \left\|\mathbb{E}_{v_k}\left[\tilde{\phi}(Zv_k)E_k v_k^\top v_k E_k \tilde{\phi}(Zv_k)^\top\right]\right\|_{\mathsf{op}} \\ &\leq \left\|\mathbb{E}_{v_k}\left[\tilde{\phi}(Zv_k)\tilde{\phi}(Zv_k)^\top\right]\right\|_{\mathsf{op}} \sup_{v_k}\left(E_k\|v_k\|_2^2\right) \\ &\leq C_1^2\left\|\mathbb{E}_{v_k}\left[\tilde{\phi}(Zv_k)\tilde{\phi}(Zv_k)^\top\right]\right\|_{\mathsf{op}} \\ &= O(1),\end{aligned} \tag{G.110}$$

where the last step is a direct consequence of Lemma G.9, applied to $\tilde{\Phi}$ instead of to Φ, and holds with probability at least $1 - 2\exp\left(-c_5 \log^2 d\right)$ over Z. For the other argument in the $\max$ in (G.109) we similarly have

$$\begin{aligned}\left\|\mathbb{E}\left[W_k^\top W_k\right]\right\|_{\mathsf{op}} &\leq \left\|\mathbb{E}_{v_k}\left[v_k E_k \tilde{\phi}(Zv_k)^\top \tilde{\phi}(Zv_k)E_k v_k^\top\right]\right\|_{\mathsf{op}} \\ &\leq \left\|\mathbb{E}_{v_k}\left[v_k v_k^\top\right]\right\|_{\mathsf{op}} \sup_{v_k}\left(E_k\left\|\tilde{\phi}(Zv_k)\right\|_2^2\right) \\ &\leq \frac{1}{d}C_1^2 d \\ &= O(1).\end{aligned} \tag{G.111}$$

Then, plugging these last two equations in (G.108) we get

$$\begin{aligned}\mathbb{P}_V\left(\left\|\tilde{\Phi}EV - \mathbb{E}_V\left[\tilde{\Phi}EV\right]\right\|_{\mathsf{op}} \geq \sqrt{p}\log d\right) &\leq (n+d)\exp\left(-\frac{p\log^2 d/2}{C_2 p + 2C_1^2\sqrt{d}\sqrt{p}\log d/3}\right) \\ &\leq 2\exp\left(-c_6\log^2 d\right),\end{aligned} \tag{G.112}$$

where we used Assumption 6.8. Then, applying a triangle inequality and using (G.105) and (G.112), we get

$$\left\|\tilde{\Phi}EV\right\|_{\mathsf{op}} = O\left(\sqrt{p}\log d + \frac{p\log d}{d}\right), \tag{G.113}$$

with probability at least $1 - 2\exp\left(-c_7\log^2 d\right)$ over Z, V. Then, since $E = I$ with probability at least $1 - 2p\exp\left(-c_3 d\right)$, using Assumption 6.8 we get the desired result.

□

Lemma G.14. *Let $z \in \mathbb{R}^{2d}$ be sampled from a distribution respecting Assumption 6.9, not necessarily with the same covariance as $\mathcal{P}_{XY}$, independent from everything else, and let $f_{\mathsf{RF}}(\hat{\theta}_{\mathsf{RF}}(\lambda), z)$ be the RF model defined in* (6.22) *with $\hat{\theta}_{\mathsf{RF}}(\lambda)$ defined in* (6.24)*, with $\lambda \geq 0$. Then, we have that*

$$\left| f_{\mathsf{RF}}(\hat{\theta}_{\mathsf{RF}}(\lambda), z) - \mu_1^2 p \frac{z^\top Z^\top}{2d}\left(\Phi\Phi^\top + n\lambda I\right)^{-1} G \right| = O\left(\frac{d^{1/4}\log d}{p^{1/4}} + \frac{\log^{3/2} d}{d^{1/8}}\right) = o(1), \tag{G.114}$$

with probability $1 - C\sqrt{d}\log^2 d/\sqrt{p} - C\log^3 d/d^{1/4}$ over Z, G, V and z, where C is an absolute constant

Proof. Let $\tilde{\phi} : \mathbb{R} \to \mathbb{R}$ be defined as $\tilde{\phi}(\cdot) := \phi(\cdot) - \mu_1(\cdot)$, and let $\tilde{\Phi} \in \mathbb{R}^{n\times p}$ be defined as the matrix containing $\tilde{\phi}\left(Vz_i\right)$ in its i-th row. Then, introducing the shorthand $\hat{G} = \left(\Phi\Phi^\top + n\lambda I\right)^{-1} G$, we can write

$$\begin{aligned} f_{\mathsf{RF}}(\hat{\theta}_{\mathsf{RF}}(\lambda), z) &= \left(\mu_1 Vz + \tilde{\phi}\left(Vz\right)\right)^\top \left(\mu_1 VZ^\top + \tilde{\Phi}^\top\right)\hat{G} \\ &= \mu_1^2 p\frac{z^\top Z^\top}{2d}\hat{G} + \mu_1^2 z^\top\left(V^\top V - \frac{p}{2d}I\right)Z^\top\hat{G} + \mu_1 z^\top V^\top\tilde{\Phi}^\top\hat{G} + \tilde{\phi}\left(Vz\right)^\top\Phi\hat{G}. \end{aligned} \tag{G.115}$$

Notice that since every entry of G is sub-Gaussian and independent by Assumption 6.9, Theorem 3.1.1 in Vershynin [2018] readily gives $\|G\|_2 = O\left(\sqrt{n}\right) = O\left(\sqrt{d}\right)$ with probability at least $1 - 2\exp(-c_1 d)$ over G. Then, conditioning on the high probability event described by Lemma G.10, we get

$$\left\|\hat{G}\right\|_2 \leq \left(\lambda_{\min}\left(\Phi\Phi^\top + n\lambda I\right)\right)^{-1}\|G\|_2 \leq \left(\lambda_{\min}\left(\Phi\Phi^\top\right)\right)^{-1}\|G\|_2 = O\left(\frac{\sqrt{d}}{p}\right), \tag{G.116}$$

with probability at least $1 - 2\exp\left(-c_2\log^2 d\right)$ over V, Z and G. We will condition on this high probability event until the end of the proof. Let's then investigate the last 3 terms on the RHS of (G.115) separately:

(i) A direct application of Theorem 5.39 of Vershynin [2012] (see their Equation 5.23) gives

$$\left\|\frac{2d}{p}V^\top V - I\right\|_{\mathsf{op}} = O\left(\sqrt{\frac{d}{p}}\right), \tag{G.117}$$

with probability at least $1 - 2\exp\left(-c_3 d\right)$ over V. Then, since z is sub-Gaussian and independent from everything else, with probability $1 - 2\exp\left(-c_4\log^2 d\right)$ over itself we

have

$$\begin{aligned}\left|\mu_1^2 z^\top \left(V^\top V - \frac{p}{2d} I\right) Z^\top \hat{G}\right| &\le \log d \left\|\left(V^\top V - \frac{p}{2d} I\right) Z^\top \hat{G}\right\|_2 \\ &\le \log d \left\|V^\top V - \frac{p}{2d} I\right\|_{\mathsf{op}} \|Z\|_{\mathsf{op}} \left\|\hat{G}\right\|_2 \\ &= O\left(\log d \sqrt{\frac{p}{d}} \sqrt{d} \frac{\sqrt{d}}{p}\right) = O\left(\frac{\sqrt{d}\log d}{\sqrt{p}}\right),\end{aligned} \tag{G.118}$$

where the third step holds with probability at least $1 - 2\exp(-c_5 d)$ over Z due to Lemma G.8.

(ii) As before, since z is sub-Gaussian and independent from everything else, with probability $1 - 2\exp\left(-c_4 \log^2 d\right)$ we have

$$\begin{aligned}\left|\mu_1 z^\top V^\top \tilde{\Phi}^\top \hat{G}\right| &\le \log d \left\|V^\top \tilde{\Phi}^\top\right\|_{\mathsf{op}} \left\|\hat{G}\right\|_2 \\ &= O\left(\log d \left(\sqrt{p}\log d + \frac{p\log d}{d}\right) \frac{\sqrt{d}}{p}\right) = O\left(\log^2 d \left(\sqrt{\frac{d}{p}} + \frac{1}{\sqrt{d}}\right)\right),\end{aligned} \tag{G.119}$$

where the second step holds because of Lemma G.13, and holds with probability at least $1 - 2\exp\left(-c_6 \log^2 d\right)$ over Z and V.

(iii) For the last term of the RHS of (G.115), its second moment in the probability space of z reads

$$\begin{aligned}\mathbb{E}_z\left[\hat{G}^\top \Phi^\top \tilde{\phi}(Vz)\, \tilde{\phi}(Vz)^\top \Phi \hat{G}\right] &\le \left\|\mathbb{E}_z\left[\tilde{\phi}(Vz)\, \tilde{\phi}(Vz)^\top\right]\right\|_{\mathsf{op}} \|\Phi\|_{\mathsf{op}}^2 \left\|\hat{G}\right\|_2^2 \\ &= O\left(\left(\log^4 d + \frac{p\log^3 d}{d^{3/2}}\right) \sqrt{d}\frac{\sqrt{d}}{p}\right) \\ &= O\left(\frac{d\log^4 d}{p} + \frac{\log^3 d}{\sqrt{d}}\right),\end{aligned} \tag{G.120}$$

where the second step follows from Lemmas G.12 and G.10, and holds with probability at least $1 - 2\exp\left(-c_7 \log^2 d\right)$ over Z and V. Then, by Markov inequality, we have that there exists a constant C_1 such that

$$\left(\tilde{\phi}(Vz)^\top \Phi \hat{G}\right)^2 < C_1 \left(\frac{d\log^4 d}{p} + \frac{\log^3 d}{\sqrt{d}}\right) t, \tag{G.121}$$

with probability at least $1 - 1/t$ over z. Setting

$$t = \min\left(\frac{\sqrt{p}}{\sqrt{d}\log^2 d}, d^{1/4}\right) = \omega(1), \tag{G.122}$$

since $p = \omega\left(d\log^4 d\right)$ by Assumption 6.8, we have

$$\left|\tilde{\phi}(Vz)^\top \Phi \hat{G}\right| = O\left(\frac{d^{1/4}\log d}{p^{1/4}} + \frac{\log^{3/2} d}{d^{1/8}}\right), \tag{G.123}$$

with probability at least $1 - \sqrt{d}\log^2 d/\sqrt{p} - \log^3 d/d^{1/4}$.

Then, plugging (i), (ii) and (iii) in (G.115) gives

$$\left| f_{\mathsf{RF}}(\hat{\theta}_{\mathsf{RF}}(\lambda), z) - \mu_1^2 p \frac{z^\top Z^\top}{2d} \hat{G} \right| = O\left(\frac{d^{1/4} \log d}{p^{1/4}} + \frac{\log^{3/2} d}{d^{1/8}} \right), \tag{G.124}$$

with probability at least $1 - c_8 \frac{\sqrt{d} \log^2 d}{\sqrt{p}} - c_8 \frac{\log^3 d}{d^{1/4}}$, which gives the desired result. □

Proof of Theorem 6.10 Let $E \in \mathbb{R}^{n \times n}$ be the matrix defined as

$$E = \Phi\Phi^\top - p\left(\mu_1^2 \frac{ZZ^\top}{2d} + \tilde{\mu}^2 I \right). \tag{G.125}$$

Note that

$$\begin{aligned} \|E\|_{\mathsf{op}} &\leq \left\| \Phi\Phi^\top - \mathbb{E}_V\left[\Phi\Phi^\top\right] \right\|_{\mathsf{op}} + \left\| \mathbb{E}_V\left[\Phi\Phi^\top\right] - p\left(\mu_1^2 \frac{ZZ^\top}{2d} + \tilde{\mu}^2 I \right) \right\|_{\mathsf{op}} \\ &= O\left(p\left(\sqrt{\frac{d}{p}} + \frac{\log^3 d}{\sqrt{d}} \right) \right), \end{aligned} \tag{G.126}$$

with probability at least $1 - 2\exp\left(-c_1 \log^2 d\right)$ over Z, V due to Lemmas G.9 and G.10. By the Woodbury matrix identity (or Hua's identity), we have

$$\begin{aligned} &\left(\Phi\Phi^\top + n\lambda I\right)^{-1} = \\ &= \left(p\left(\mu_1^2 \frac{ZZ^\top}{2d} + \tilde{\mu}^2 I \right) + E + n\lambda I \right)^{-1} \\ &= \left(\mu_1^2 p \frac{ZZ^\top}{2d} + \left(\tilde{\mu}^2 p + n\lambda\right) I + E \right)^{-1} \\ &= \left(\mu_1^2 p \frac{ZZ^\top}{2d} + \left(\tilde{\mu}^2 p + n\lambda\right) I \right)^{-1} \\ &\quad - \left(\mu_1^2 p \frac{ZZ^\top}{2d} + \left(\tilde{\mu}^2 p + n\lambda\right) I \right)^{-1} E \left(\mu_1^2 p \frac{ZZ^\top}{2d} + \left(\tilde{\mu}^2 p + n\lambda\right) I + E \right)^{-1}, \end{aligned} \tag{G.127}$$

which gives

$$\begin{aligned} &\left| \mu_1^2 p \frac{z^\top Z^\top}{2d} \left(\Phi\Phi^\top + n\lambda I\right)^{-1} G - \mu_1^2 p \frac{z^\top Z^\top}{2d} \left(\mu_1^2 p \frac{ZZ^\top}{2d} + \left(\tilde{\mu}^2 p + n\lambda\right) I \right)^{-1} G \right| \leq \\ &\leq \left| \mu_1^2 p \frac{z^\top Z^\top}{2d} \left(\mu_1^2 p \frac{ZZ^\top}{2d} + \left(\tilde{\mu}^2 p + n\lambda\right) I \right)^{-1} E \left(\mu_1^2 p \frac{ZZ^\top}{2d} + \left(\tilde{\mu}^2 p + n\lambda\right) I + E \right)^{-1} G \right| \\ &\leq \log d \left\| \frac{p}{2d} Z^\top \left(\mu_1^2 p \frac{ZZ^\top}{2d} + \left(\tilde{\mu}^2 p + n\lambda\right) I \right)^{-1} E \left(\mu_1^2 p \frac{ZZ^\top}{2d} + \left(\tilde{\mu}^2 p + n\lambda\right) I + E \right)^{-1} G \right\|_2 \\ &\leq \log d \, \frac{p}{2d} \|Z\|_{\mathsf{op}} \frac{1}{\tilde{\mu}^2 p} \|E\|_{\mathsf{op}} \frac{1}{\lambda_{\min}\left(\Phi\Phi^\top\right)} \|G\|_2 \\ &= O\left(\log d \, \frac{p}{d} \sqrt{d} \, \frac{1}{p} \, p \left(\sqrt{\frac{d}{p}} + \frac{\log^3 d}{\sqrt{d}} \right) \frac{1}{p} \sqrt{d} \right) \\ &= O\left(\log d \sqrt{\frac{d}{p}} + \frac{\log^4 d}{\sqrt{d}} \right). \end{aligned} \tag{G.128}$$

Here, the second step holds with probability at least $1 - 2\exp\left(-c_2 \log^2 d\right)$ since z is sub-Gaussian and independent from everything else; the fourth step is a consequence of Lemma G.8, (G.126), Lemma G.10, and $\|G\|_2 = O\left(\sqrt{d}\right)$ (see the argument prior to (G.116)), and as a whole holds with probability $1 - 2\exp\left(-c_3 \log^2 d\right)$ over Z, G, and V.

Note that the second term in the LHS of (G.128) can be written as

$$\begin{aligned}\mu_1^2 p \frac{z^\top Z^\top}{2d}\left(\mu_1^2 p \frac{ZZ^\top}{2d} + \left(\tilde{\mu}^2 p + n\lambda\right) I\right)^{-1} G &= z^\top Z^\top \left(ZZ^\top + n\left(\frac{2\tilde{\mu}^2 d}{\mu_1^2 n} + \frac{2d}{\mu_1^2 p}\lambda\right) I\right)^{-1} G \\ &= z^\top \left(Z^\top Z + n\left(\frac{2\tilde{\mu}^2 d}{\mu_1^2 n} + \frac{2d}{\mu_1^2 p}\lambda\right) I\right)^{-1} Z^\top G \\ &= f_{\mathsf{LR}}(\hat{\theta}_{\mathsf{LR}}(\tilde{\lambda}), z),\end{aligned} \tag{G.129}$$

where the second line is due to the classical identity $A^\top \left(AA^\top + \kappa I\right)^{-1} = \left(A^\top A + \kappa I\right)^{-1} A^\top$, and the third line uses the definition in (6.5), with $\hat{\theta}_{\mathsf{LR}}(\tilde{\lambda})$ defined in (6.8) and

$$\tilde{\lambda} = \frac{2\tilde{\mu}^2 d}{\mu_1^2 n} + \frac{2d}{\mu_1^2 p}\lambda. \tag{G.130}$$

Furthermore, the first term of the LHS of (G.128) satisfies

$$\left|\mu_1^2 p \frac{z^\top Z^\top}{2d}\left(\Phi\Phi^\top + n\lambda I\right)^{-1} G - f_{\mathsf{RF}}(\hat{\theta}_{\mathsf{RF}}(\lambda), z)\right| = O\left(\frac{d^{1/4}\log d}{p^{1/4}} + \frac{\log^{3/2} d}{d^{1/8}}\right), \tag{G.131}$$

with probability $1 - C_1\sqrt{d}\log^2 d/\sqrt{p} - C_1 \log^3 d/d^{1/4}$ over Z, G, V and z, due to Lemma G.14. Thus, an application of the triangle inequality together with (G.128), (G.129) and (G.131) gives the desired result. □

G.3 Features Composed as $z = x + y$

In this work, we consider the data to be composed as $z = [x^\top, y^\top]^\top$. However, some of the high level results we obtain can be recovered for settings where the features are composed differently. As an example, let us consider the model

$$z = x + y, \qquad g = x^\top \theta^* + \varepsilon. \tag{G.132}$$

Then, according to (6.3), we have

$$\mathcal{C}(\hat{\theta}) = \mathrm{Cov}\left((\tilde{x} + y)^\top \hat{\theta}, x^\top \theta^*\right) = \hat{\theta}^\top \Sigma_{yx} \theta^*, \tag{G.133}$$

which provides a quantity that could be studied again via the analysis in Han and Xu [2023]. In this new setup, we remark that the full data covariance takes the form

$$\Sigma_{zz} = \Sigma_{xx} + \Sigma_{yy} + \Sigma_{xy} + \Sigma_{yx}. \tag{G.134}$$

In a nutshell, we expect the analysis for this setting to provide a qualitative behavior similar to that unveiled by the case $z = [x^\top, y^\top]^\top$. In fact, the experiments on Color-MNIST (which does not strictly follow the model $z = [x^\top, y^\top]^\top$, as the color overlaps with the core feature pattern as in the model $z = x + y$) suggest that our conclusions hold beyond the setting of "orthogonal features". We further remark that in the setting $z = [x^\top, y^\top]^\top$ the optimal solution $\hat{\theta} = \theta^*$ gives $\mathcal{C} = 0$, while this is not necessarily the case in the setting $z = x + y$.

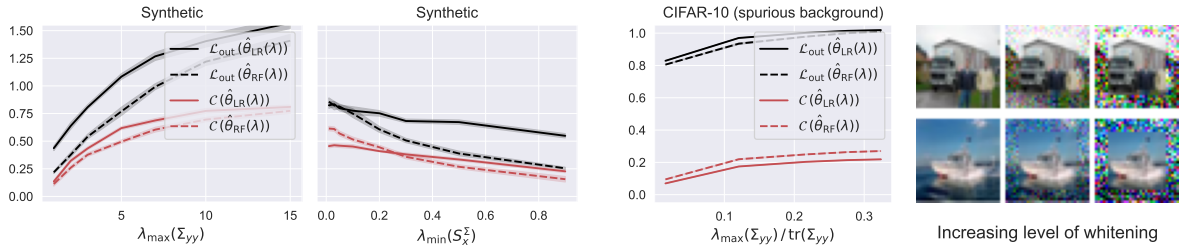


Figure G.1: Out-of-distribution test loss $\mathcal{L}(\hat{\theta}_{\text{LR/RF}}(\lambda))$ (black) and spurious correlations $\mathcal{C}(\hat{\theta}_{\text{LR/RF}}(\lambda)$ (red) as a function of $\lambda_{\max}(\Sigma_{yy})$ (first panel) and $\lambda_{\min}\left(S_x^{\Sigma}\right)$ (second panel) on a Gaussian synthetic dataset, and for the CIFAR-10 experiment (third panel). We consider the same set-up as Figures 6.3 and 6.4, for Gaussian and CIFAR-10 data, respectively.

G.4 Proof of (6.4): connection with out-of-distribution loss

Let $\tilde{x}$ and $[x^\top, y^\top]^\top$ be sampled independently from $\mathcal{P}_X$ and $\mathcal{P}_{XY}$ respectively. For simplicity, assume that $\mathbb{E}\left[f(\hat{\theta}, [\tilde{x}^\top, y])^2\right] = \mathbb{E}\left[f_x^*(\tilde{x})^2\right] = 1$ and $\mathbb{E}\left[f(\hat{\theta}, [\tilde{x}^\top, y])\right] = \mathbb{E}\left[f_x^*(\tilde{x})\right] = 0$. Thus, for the quadratic loss, we readily get

$$\begin{aligned}\mathbb{E}_{\tilde{x},y}\left[\left(f(\hat{\theta}, [\tilde{x}^\top, y]) - f_x^*(\tilde{x})\right)^2\right] &= 2 - 2\mathbb{E}_{\tilde{x},y}\left[f(\hat{\theta}, [\tilde{x}^\top, y]) f_x^*(\tilde{x})\right] \\ &= 2 - 2\text{Cov}\left(f(\hat{\theta}, [\tilde{x}^\top, y]), f_x^*(\tilde{x})\right).\end{aligned} \tag{G.135}$$

Denoting with S the covariance matrix of the three random variables $f(\hat{\theta}, [\tilde{x}^\top, y])$, $f_x^*(\tilde{x})$, and $f_x^*(x)$, we have

$$S = \begin{pmatrix} 1 & \rho & \mathcal{C} \\ \rho & 1 & 0 \\ \mathcal{C} & 0 & 1 \end{pmatrix}, \tag{G.136}$$

where we introduced $\mathcal{C} = \text{Cov}\left(f(\hat{\theta}, [\tilde{x}^\top, y]), f_x^*(x)\right)$ and $\rho = \text{Cov}\left(f(\hat{\theta}, [\tilde{x}^\top, y]), f_x^*(\tilde{x})\right)$. Since S is p.s.d., its determinant has to be non-negative, hence

$$1 - \rho^2 - \mathcal{C}^2 \geq 0, \tag{G.137}$$

which, when plugged in (G.135), gives (6.4).

This bound suggests the close connection between $\mathcal{C}$ and the out-of-distribution test loss. In Figure G.1, we repeat the same experiments of Figures 6.3-6.4 and report in black the out-of-distribution test loss. The plots clearly show that $\mathcal{C}$ and the out-of-distribution test loss follow a similar trend, for both linear regression and random features.

G.5 Experimental details

All the plots in the figures report the average over 10 independent trials, with a shaded area describing a confidence interval of 1 standard deviation. For the Gaussian and Color-MNIST datasets, every iteration involves re-generating (or re-coloring) the data, while for the CIFAR-10 dataset the randomness comes from the model and the training algorithm.

Synthetic Gaussian data generation. This follows the same model across all the numerical experiments presented in the chapter. In particular, we fix $d = 400$ and set $\Sigma_{xx} = I$. Σ_{yy} is a diagonal matrix, such that its first entry equals $\lambda_{\max}(\Sigma_{yy})$ and all the other entries equal $(d - \lambda_{\max}(\Sigma_{yy}))/(d-1)$. In this way, $\operatorname{tr}(\Sigma) = 2d$. Then, we set the off-diagonal blocks Σ_{xy} and Σ_{yx} to the same diagonal matrix, so that

$$\Sigma_{xy} = \Sigma_{yx} = (\Sigma_{yy} - \beta I)^{1/2}, \tag{G.138}$$

which implies that the Schur complement $S_x^\Sigma = \beta I$ and, therefore, $\lambda_{\min}\left(S_x^\Sigma\right) = \beta$. To conclude, we set the ground truth $\theta_x^* = e_1$, *i.e.*, the first element of the canonical basis in $\mathbb{R}^d$. This design choice is motivated by our interest in capturing the role of $\lambda_{\max}(\Sigma_{yy})$ and to have an easy control on the Schur complement S_x^Σ (which is therefore chosen to be proportional to the identity).

Unless differently stated in the figure, $n = 2000$, $\lambda_{\max}(\Sigma_{yy}) = 2$, $\beta = 0.5$, and $\lambda = 1$. Furthermore, to generate the labels, we add an independent noise with variance $\sigma^2 = 0.25$, and we subtract this quantity from the test loss, so that the optimal predictor θ^* has a test loss equal to 0.

When we use an RF model, on every dataset, unless differently stated in the figure, we use $\tanh$ as activation function, with $p = 20000$ neurons.

Binary color MNIST. This dataset is graphically shown in Figure 6.1. To generate it, we take a subset of the MNIST training dataset ($n = 1000$ samples as default, unless differently specified) made only of zeros and ones. Then, for every training image, we color the white portion in blue (red) with probability $(1 + \alpha)/2$ if the digit is a zero (one), and red (blue) otherwise. For the test set, we proceed in the same way, but setting $\alpha = 0$, to make the core feature (the digit) effectively independent from the spurious one (the color). For all the experiments, we set $\beta = 1 - \alpha^2 = 0.25$.

CIFAR-10. For the experiments on CIFAR-10, we implicitly suppose that the middle 22×22 square contains the core, predictive feature x. Thus, we sum to all the channels of the outer region white noise with increasing variance, and we later clamp the pixels to ensure their value is between 0 and 1. Increasing the variance of the noise, this progressively makes the outer portion being dominated by random noise, thus reducing its value of $\lambda_{\max}(\Sigma_{yy}) / \operatorname{tr}(\Sigma_{yy})$ when estimating the covariance on the perturbed training set. At test time, we take the images from the CIFAR-10 test set, and we add the same level of noise. To compute $\mathcal{C}$, we create an out-of-distribution dataset where the core features are randomly permuted across different backgrounds. We always consider the subset of boats and trucks, which contains $n = 10000$ images.

2-layer neural network. In the experiments shown in Figure 6.5, we consider a 2-layer neural network trained with gradient descent and quadratic loss on the Color-MNIST and CIFAR-10 datasets. For both datasets, we train for 1000 epochs, with learning rate 0.003, and batch size 1000.

G.5.1 Additional Experiments

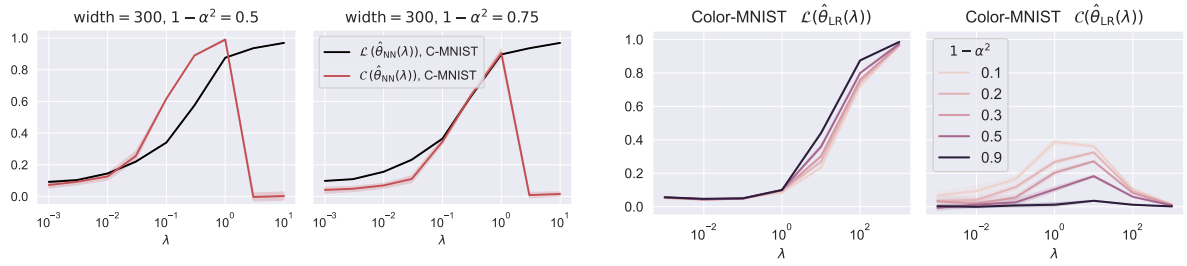


Figure G.2: Test loss $\mathcal{L}(\hat{\theta}_{\mathrm{NN/LR}}(\lambda))$ (black) and spurious correlations $\mathcal{C}(\hat{\theta}_{\mathrm{NN/LR}}(\lambda)$ (red) as a function of λ. *First and second panel:* 2-layer fully connected ReLU network, trained on the multi-class color(C)-MNIST, for two different values of α. *Third and fourth panel:* Same setup as in Figure 6.2, for the binary C-MNIST dataset, with multiple values of α.

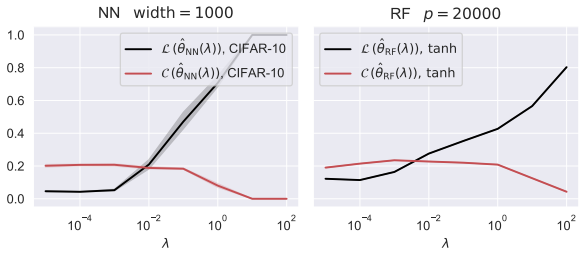


Figure G.3: Test loss $\mathcal{L}(\hat{\theta}_{\mathrm{NN/RF}}(\lambda))$ (black) and spurious correlations $\mathcal{C}(\hat{\theta}_{\mathrm{NN/RF}}(\lambda)$ (red) as a function of λ. *Left:* 2-layer fully connected ReLU network, trained on the Corrupted CIFAR-10 dataset (boats and trucks, with added textures "brightness" and "glass blur"). *Right:* RF model with $\tanh$.

In the left two panels of Figure G.2 we consider Color-MNIST including all 10 digits. We train a 2-layer network on all classes with one-hot encoding and MSE loss. Odd (even) digits are red (blue) with probability $(1+\alpha)/2$, and blue (red) with probability $(1-\alpha)/2$. To compute $\mathcal{C}$, at test time we consider the parity of the logit with the highest value, with respect to the color of the image. The first and the second panel correspond to two values of α-s and follow a similar profile as Figure 6.5 (left), showing the same qualitative behavior of $\mathcal{L}$ and $\mathcal{C}$ with respect to λ for the full Color-MNIST dataset (i.e., in the multi-class setting). In the third and fourth panel of Figure G.2 we repeat the experiment in Figure 6.2 (right) for multiple values of α, reporting $\mathcal{L}$ and $\mathcal{C}$ with respect to λ for linear regression. The curves behave as expected: for any value of λ, as α decreases, $\mathcal{C}$ decreases. Furthermore, the (in-distribution) test loss decreases as α increases, in agreement with our discussion at the end of Section 6.5.

In Figure G.3 we train a 2-layer network and a random feature model on the classes "trucks" and "boats" from the Corrupted CIFAR-10 dataset (used in the context of spurious correlations in Liu et al. [2023a]), enforcing a correlation in the training set with respectively the textures "brightness" and "glass blur", with a correlation $\alpha = 0.95$. In both panels, we see a mild increase in $\mathcal{C}$ as λ initially increases, until the later decrease predicted by Proposition 6.4. The profiles are also qualitatively similar to the ones of Figure 6.5 (left).

APPENDIX H

Appendix for Chapter 7

H.1 Proofs for random features

In this section, we provide the proofs for our results on the random features model. Thus, we will drop the sub-script "RF" in all the quantities of this section, for the sake of a cleaner notation. We consider a single textual data-point $X = [x_1, x_2, ..., x_N]^\top \in \mathbb{R}^{N\times d}$ that satisfies Assumption 7.1. We consider the random features model defined in (7.1), i.e.,

$$\varphi(X) = \phi(V \operatorname{flat}(X)), \tag{H.1}$$

where $V_{i,j} \sim_{\text{i.i.d.}} \mathcal{N}(0, 1/D)$, $D = Nd$ and ϕ is the activation function, applied component-wise to the pre-activations $V \operatorname{flat}(X)$.

Further in the section, we will investigate the generalization capabilities of the RF model $f(\cdot, \theta) := \varphi(\cdot)^\top \theta$ on token modification. We use the notation $(\mathcal{X}, \mathcal{Y})$ to indicate the original labelled training dataset, with $\mathcal{X} = (X_1, \dots, X_n)$, $X_i \in \mathbb{R}^{N\times d}$ and $\mathcal{Y} = [y_1, \dots, y_n]^\top \in \{-1, +1\}^n$. We will use the short-hand $\Phi := [\varphi(X_1), ...\varphi(X_n)]^\top \in \mathbb{R}^{n\times p}$ for the feature matrix and $K := \Phi\Phi^\top$ for the kernel. According to (7.12), minimizing the quadratic loss over this dataset returns the parameters

$$\theta^* = \theta_0 + \Phi^+(\mathcal{Y} - \Phi\theta_0). \tag{H.2}$$

We consider the test error on the modified sample $X^i(\Delta)$ given by (see also (7.16))

$$\text{Err}(X^i(\Delta), \theta) := \Big(f(X^i(\Delta), \theta) - y_\Delta\Big)^2. \tag{H.3}$$

We will investigate this quantity for both the fine-tuned and the re-trained model. In particular, the new solution obtained after fine-tuning over the sample (X, y) gives

$$\theta_f^* = \theta^* + \frac{\varphi(X)}{\|\varphi(X)\|_2^2}\left(y - \varphi(X)^\top \theta^*\right), \tag{H.4}$$

while retraining the model with initialization θ_0 on the new dataset $(\mathcal{X}_r, \mathcal{Y}_r)$ returns

$$\theta_r^* = \theta_0 + \Phi_r^+(\mathcal{Y}_r - \Phi_r\theta_0), \tag{H.5}$$

where we denote by $\Phi_r := [\varphi(X_1), ...\varphi(X_n), \varphi(X)]^\top \in \mathbb{R}^{(n+1)\times p}$ the new feature matrix. The corresponding kernel is denoted by $K_r = \Phi_r\Phi_r^\top$.

The outline of this section is the following:

1. We use Lemma H.1 to upper bound the numerator of the word sensitivity $\mathcal{S}(X)$ defined in (7.4), which readily allows to prove the desired result Theorem 7.2.

2. We prove Theorem 7.6, where we lower bound $\text{Err}(X^i(\Delta), \theta_f^*)$ for the fine-tuned solution θ_f^* as a function of γ.

3. We prove Theorem 7.7, where we lower bound $\text{Err}(X^i(\Delta), \theta_r^*)$ for the retrained solution θ_r^* as a function of γ. This result requires additional assumptions and analysis:

   - We report in our notation Lemma 4.1 from Bombari and Mondelli [2024a], which defines the *feature alignment* $\mathcal{F}(X, X^i(\Delta))$ between the two samples X and $X^i(\Delta)$.
   - In Lemma H.2, we exploit the additional assumptions to prove a lower bound on the smallest eigenvalue of the kernel $\lambda_{\min}(K_r) = \Omega(k)$.
   - In Lemma H.3, we show that $|\mathcal{F}(X, X^i(\Delta)) - 1| = O\left(1/\sqrt{N}\right)$ with high probability. This step is useful as here this term assumes the role previously taken by $\mathcal{S}(X)$.

Lemma H.1. *Let $\varphi(X)$ be the random feature map defined in* (7.1)*, with ϕ Lipschitz and $k = \Omega(D)$. Let $\Delta \in \mathbb{R}^d$ be such that $\|\Delta\|_2 \leq \sqrt{d}$. Then, for every $i \in [N]$, we have*

$$\left\|\varphi(X^i(\Delta)) - \varphi(X)\right\|_2 = O\left(\sqrt{\frac{k}{N}}\right), \tag{H.6}$$

with probability at least $1 - \exp(-cD)$ over V.

Proof. Let's condition on the event

$$\|V\|_{\text{op}} = O\left(\sqrt{\frac{D+k}{D}}\right), \tag{H.7}$$

which happens with probability at least $1 - \exp(-c_1 D)$ over V, by Theorem 4.4.5 of Vershynin [2018]. Thus, for every i, we have

$$\begin{aligned}
\left\|\varphi(X^i(\Delta)) - \varphi(X)\right\|_2 &= \left\|\phi(V \,\text{flat}\,(X^i(\Delta))) - \phi(V \,\text{flat}\,(X))\right\|_2 \\
&\leq M \left\|V(\text{flat}\,(X^i(\Delta)) - \text{flat}\,(X))\right\|_2 \\
&\leq M \|V\|_{\text{op}} \|\Delta\|_2 \\
&\leq C_1 \sqrt{\frac{D+k}{D}} \sqrt{d} \\
&= C_1 \sqrt{\frac{D+k}{N}} \\
&\leq C_2 \sqrt{\frac{k}{N}},
\end{aligned} \tag{H.8}$$

where the first inequality comes from the Lipschitz continuity of ϕ, and the last step is a consequence of $k = \Omega(D)$. □

Theorem 7.2 Let $\varphi(X)$ be the random feature map defined in (7.1), where ϕ is Lipschitz and not identically 0. Let $X \in \mathbb{R}^{N\times d}$ be a generic input sample s.t. Assumption 7.1 holds, and assume $k = \Omega(D)$. Let $\mathcal{S}(X)$ denote the the word sensitivity defined in (7.4). Then, we have

$$\mathcal{S}(X) = O\left(1/\sqrt{N}\right) = o(1), \tag{H.9}$$

with probability at least $1 - \exp(-cD)$ over V.

Proof. As ϕ is Lipschitz and non-0, we can apply the result in Lemma C.3 of Bombari and Mondelli [2024a], getting

$$\|\varphi(X)\|_2 = \Theta(\sqrt{k}), \tag{H.10}$$

with probability at least $1 - \exp(-c_1 D)$ over V. Thus, the thesis readily follows from Lemma H.1. □

Proof of Remark 7.4. The only difference with respect to the argument for Theorem 7.2 is in (H.8). Now, Δ is replaced by a new set of m perturbations $\Delta_1, \dots, \Delta_m$. Thus, the modified context takes the form

$$X(\Delta_1, \dots, \Delta_m) = X + \sum_{j=1}^{m} e_{i_j} \Delta_j^\top, \tag{H.11}$$

where $\{e_{i_j}\}_{j\in[m]}$ represent different elements of the canonical basis. Thus, we have

$$\|\mathrm{flat}\,(X(\Delta_1, \dots, \Delta_m)) - \mathrm{flat}\,(X)\|_2 = \left\|\mathrm{flat}\left(\sum_{j=1}^{m} e_{i_j}\Delta_j^\top\right)\right\|_2 = \left\|\sum_{j=1}^{m} e_{i_j}\Delta_j^\top\right\|_F. \tag{H.12}$$

Since the e_{i_j}'s are all distinct (as we are modifying m different words), we obtain

$$\left\|\sum_{j=1}^{m} e_{i_j}\Delta_j^\top\right\|_F^2 = \sum_{j=1}^{m} \|\Delta_j\|_2^2 \le md, \tag{H.13}$$

where in the last step we use that $\|\Delta_j\|_2 \le \sqrt{d}$ for all $j \in [m]$.

This allows to replace the $\sqrt{d}$ in the fourth line of (H.8) with $\sqrt{md}$, increasing the final bound in the statement of Lemma H.1 by a factor $\sqrt{m}$. Thus, the upper bound on the sensitivity is given by $\sqrt{m/N} = o(1)$, which concludes the proof. □

Theorem 7.6 Let $f(\cdot, \theta_f^*)$ be the RF model fine-tuned on the sample (X, y), where ϕ in (7.1) is Lipschitz and not identically 0, $X \in \mathbb{R}^{N\times d}$ is a generic sample s.t. Assumption 7.1 holds and θ_f^* is defined in (7.14). Assume $k = \Omega(D)$, $|f(X, \theta^*)| = O(1)$, and that Assumption 7.5 holds with $\gamma \in [0, 2)$. Let $\mathrm{Err}(X^i(\Delta), \theta_f^*)$ be the test error of the model on the sample $X^i(\Delta)$ defined in (7.16). Then, for any Δ such that $\|\Delta\|_2 \le \sqrt{d}$ and any $i \in [N]$, we have

$$\mathrm{Err}(X^i(\Delta), \theta_f^*) > (2-\gamma)^2 - O\left(1/\sqrt{N}\right), \tag{H.14}$$

with probability at least $1 - \exp(-cD)$ over V.

Proof. We have

$$f(X^i(\Delta),\theta_f^*)=\varphi(X^i(\Delta))^\top\theta_f^*=\varphi(X^i(\Delta))^\top\theta^*+\frac{\varphi(X^i(\Delta))^\top\varphi(X)}{\|\varphi(X)\|_2^2}\left(y-\varphi(X)^\top\theta^*\right), \tag{H.15}$$

where the second step is justified by (7.14). As $\left|\varphi(X)^\top\theta^*-\varphi(X^i(\Delta))^\top\theta^*\right|=|f(X,\theta^*)-f(X^i(\Delta),\theta^*)|\leq\gamma$ by Assumption 7.5, we can write

$$\begin{aligned}\left|f(X^i(\Delta),\theta_f^*)-y\right|&=\left|f(X^i(\Delta),\theta^*)-f(X,\theta^*)+\left(\frac{\varphi(X^i(\Delta))^\top\varphi(X)}{\|\varphi(X)\|_2^2}-1\right)(y-f(X,\theta^*))\right|\\&\leq\gamma+\left|\frac{\varphi(X^i(\Delta))^\top\varphi(X)}{\|\varphi(X)\|_2^2}-1\right||y-f(X,\theta^*)|.\end{aligned} \tag{H.16}$$

By Cauchy-Schwartz inequality, we have

$$\begin{aligned}\left|\frac{\varphi(X^i(\Delta))^\top\varphi(X)}{\|\varphi(X)\|_2^2}-1\right|&=\left|\frac{\left(\varphi(X^i(\Delta))-\varphi(X)\right)^\top\varphi(X)}{\|\varphi(X)\|_2^2}\right|\\&\leq\frac{\|\varphi(X^i(\Delta))-\varphi(X)\|_2\,\|\varphi(X)\|_2}{\|\varphi(X)\|_2^2}\\&\leq\mathcal{S}(X).\end{aligned} \tag{H.17}$$

By Theorem 7.2, we have that $\mathcal{S}(X)=O\left(1/\sqrt{N}\right)$ with probability at least $1-\exp(-cD)$ over V. Conditioning on such high probability event, we can write

$$\begin{aligned}\left|f(X^i(\Delta),\theta_f^*)-y_\Delta\right|&\geq|y_\Delta-y|-\left|f(X^i(\Delta),\theta_f^*)-y\right|\\&\geq|y_\Delta-y|-\gamma-\mathcal{S}(X)\,|y-f(X,\theta^*)|\\&=2-\gamma-S(X)\,|y-f(X,\theta^*)|\\&=2-\gamma-O\left(1/\sqrt{N}\right),\end{aligned} \tag{H.18}$$

where the third step is a consequence of $|y|=|y_\Delta|=1$, with $y=-y_\Delta$, and the fourth step comes from $|f(X,\theta^*)|=O(1)$. Thus, we can conclude

$$\text{Err}(X^i(\Delta),\theta_f^*)=\left(f(X^i(\Delta),\theta_f^*)-y_\Delta\right)^2\geq\left(2-\gamma-O\left(1/\sqrt{N}\right)\right)^2>(2-\gamma)^2-O\left(1/\sqrt{N}\right). \tag{H.19}$$

□

Lemma 4.1 Bombari and Mondelli [2024a] Let the kernel $K_r\in\mathbb{R}^{(n+1)\times(n+1)}$ be invertible, and let $P_\Phi\in\mathbb{R}^{k\times k}$ be the projector over $\mathrm{Span}\{\mathrm{rows}(\Phi)\}$. Let us denote by

$$\mathcal{F}(X,X^i(\Delta)):=\frac{\varphi(X^i(\Delta))^\top P_\Phi^\perp\varphi(X)}{\left\|P_\Phi^\perp\varphi(X)\right\|_2^2} \tag{H.20}$$

the *feature alignment* between X and $X^i(\Delta)$. Then, we have

$$f(X^i(\Delta),\theta_r^*)-f(X^i(\Delta),\theta^*)=\mathcal{F}(X,X^i(\Delta))\left(f(X,\theta_r^*)-f(X,\theta^*)\right). \tag{H.21}$$

Notice that

$$\left\|P_\Phi^\perp\varphi(X)\right\|_2^2\geq\lambda_{\min}\left(K_r\right)>0 \tag{H.22}$$

is directly implied by the invertibility of K_r, as shown in Lemma B.1 from Bombari and Mondelli [2024a].

Lemma H.2. *Let ϕ be a non-linear, Lipschitz function. Let all the training data in $\mathcal{X}_r$ be sampled i.i.d. according to a distribution $\mathcal{P}_X$ s.t. $\mathbb{E}_{X\sim\mathcal{P}_X}[X] = 0$, Assumption 7.1 holds, and the Lipschitz concentration property is satisfied. Let $n\log^3 n = o(k)$ and $n\log^4 n = o(D^2)$. Then, we have*

$$\lambda_{\min}(K_r) = \Omega(k), \tag{H.23}$$

with probability at least $1 - \exp\left(-c\log^2 n\right)$ over V and $\mathcal{X}_r$.

Proof. The desired result follows from Lemma D.2 in Bombari and Mondelli [2024a]. Notice that for their argument to go through, they need their Lemma D.1 to hold, which requires our assumptions on the data distribution $\mathcal{P}_X$, and Assumption 7.1. They further require the scalings $n = o(D^2/\log^4 D)$ and $n\log^4 n = o(D^2)$, which are both given by our assumption $n\log^4 n = o(D^2)$. Finally, in their Lemma D.2 they require the activation function ϕ to be Lipschitz and non-linear, and the over-parameterized setting $n\log^3 n = o(k)$. □

Lemma H.3. *Let ϕ be a non-linear, Lipschitz function. Let all the training data in $\mathcal{X}_r$ be sampled i.i.d. according to a distribution s.t. $\mathbb{E}_{X\sim\mathcal{P}_X}[X] = 0$, Assumption 7.1 holds, and the Lipschitz concentration property is satisfied. Let $n\log^3 n = o(k)$, $n\log^4 n = o(D^2)$ and $k = \Omega(D)$. Let $\mathcal{F}(X, X^i(\Delta))$ be defined as in* (H.20)*. Then, we have*

$$\left|\mathcal{F}(X, X^i(\Delta)) - 1\right| = O\left(\frac{1}{\sqrt{N}}\right), \tag{H.24}$$

with probability at least $1 - \exp\left(-c\log^2 n\right)$ over $\mathcal{X}_r$ and V.

Proof. By Cauchy-Schwartz inequality, we have

$$\begin{aligned}
\left|\mathcal{F}(X, X^i(\Delta)) - 1\right| &= \left|\frac{(\varphi(X^i(\Delta)) - \varphi(X))^\top P_\Phi^\perp \varphi(X)}{\left\|P_\Phi^\perp \varphi(X)\right\|_2^2}\right| \\
&\le \frac{\|\varphi(X^i(\Delta)) - \varphi(X)\|_2}{\|\varphi(X)\|_2} \frac{\|\varphi(X)\|_2}{\left\|P_\Phi^\perp \varphi(X)\right\|_2} \\
&\le \mathcal{S}(X)\frac{\|\varphi(X)\|_2}{\sqrt{\lambda_{\min}(K_r)}},
\end{aligned} \tag{H.25}$$

where the last step is a consequence of (H.22). As ϕ is Lipschitz and non-0, we can apply the result in Lemma C.3 of Bombari and Mondelli [2024a], getting

$$\|\varphi(X)\|_2 = \Theta(\sqrt{k}), \tag{H.26}$$

with probability at least $1 - \exp(-c_1 D)$ over V. Due to Lemma H.2, we can also write

$$\lambda_{\min}(K_r) = \Omega(k), \tag{H.27}$$

with probability at least $1 - \exp\left(-c_2\log^2 n\right)$ over V and $\mathcal{X}_r$. Thus, (H.25) promptly gives

$$\left|\mathcal{F}(X, X^i(\Delta)) - 1\right| \le \mathcal{S}(X)\frac{\|\varphi(X)\|_2}{\sqrt{\lambda_{\min}(K_r)}} = O\left(\mathcal{S}(X)\right) = O\left(\frac{1}{\sqrt{N}}\right), \tag{H.28}$$

where the last step comes from Theorem 7.2, and holds with probability at least $1 - \exp(-c_3 D)$. This gives the desired result. □

Theorem 7.7 Let $f(\cdot, \theta_r^*)$ be the RF model re-trained on the dataset $(\mathcal{X}_r, \mathcal{Y}_r)$ that contains the pair (X, y), thus with θ_r^* defined in (7.15). Let ϕ in (7.1) be a non-linear, Lipschitz function. Assume the training data to be sampled i.i.d. from a distribution $\mathcal{P}_X$ s.t. $\mathbb{E}_{X\sim\mathcal{P}_X}[X] = 0$, Assumption 7.1 holds, and the Lipschitz concentration property is satisfied. Let $n\log^3 n = o(k)$, $n\log^4 n = o(D^2)$ and $k = \Omega(D)$. Assume that $|f(X, \theta^*)| = O(1)$ and that Assumption 7.5 holds with $\gamma \in [0, 2)$. Let $\mathsf{Err}(X^i(\Delta^*), \theta_r^*)$ be the test error of the model on the sample $X^i(\Delta)$ defined in (7.16). Then, for any Δ s.t. $\|\Delta\|_2 \leq \sqrt{d}$ and any $i \in [N]$, we have

$$\mathsf{Err}(X^i(\Delta^*), \theta_r^*) > (2-\gamma)^2 - O\left(1/\sqrt{N}\right), \tag{H.29}$$

with probability at least $1 - \exp\left(-c\log^2 n\right)$ over V and $\mathcal{X}_r$.

Proof. Let's condition on K_r being invertible, which by Lemma H.2 happens with probability at least $1 - \exp\left(-c_1\log^2 n\right)$ over V and $\mathcal{X}_r$. Then, by (H.21), we have

$$f(X^i(\Delta), \theta_r^*) - f(X^i(\Delta), \theta^*) = \mathcal{F}(X, X^i(\Delta))\left(y - f(X, \theta^*)\right), \tag{H.30}$$

since $f(\cdot, \theta_r^*)$ fully interpolates the training data, thus giving $f(X, \theta_r^*) = y$. As $\left|\varphi(X)^\top\theta^* - \varphi(X^i(\Delta))^\top\theta^*\right| = |f(X, \theta^*) - f(X^i(\Delta), \theta^*)| \leq \gamma$ by Assumption 7.5, we can write

$$\begin{aligned}\left|f(X^i(\Delta), \theta_r^*) - y\right| &= \left|f(X^i(\Delta), \theta^*) - f(X, \theta^*) + \left(\mathcal{F}(X, X^i(\Delta)) - 1\right)\left(y - f(X, \theta^*)\right)\right| \\ &\leq \gamma + \left|\mathcal{F}(X, X^i(\Delta)) - 1\right| |y - f(X, \theta^*)|.\end{aligned} \tag{H.31}$$

By Lemma H.3 we have that $|\mathcal{F}(X, X^i(\Delta)) - 1| = O\left(1/\sqrt{N}\right)$ with probability at least $1 - \exp\left(-c\log^2 n\right)$ over $\mathcal{X}_r$ and V. Conditioning on such high probability event, we can write

$$\begin{aligned}\left|f(X^i(\Delta), \theta_r^*) - y_\Delta\right| &\geq |y_\Delta - y| - \left|f(X^i(\Delta), \theta_r^*) - y\right| \\ &\geq |y_\Delta - y| - \gamma - \left|\mathcal{F}(X, X^i(\Delta)) - 1\right| |y - f(X, \theta^*)| \\ &= 2 - \gamma - \left|\mathcal{F}(X, X^i(\Delta)) - 1\right| |y - f(X, \theta^*)| \\ &= 2 - \gamma - O\left(1/\sqrt{N}\right),\end{aligned} \tag{H.32}$$

where the third step is a consequence of $|y| = |y_\Delta| = 1$, with $y = -y_\Delta$, and the fourth step comes from $|f(X, \theta^*)| = O(1)$. Thus, we can conclude

$$\mathsf{Err}(X^i(\Delta), \theta_r^*) = \left(f(X^i(\Delta), \theta_r^*) - y_\Delta\right)^2 \geq \left(2 - \gamma - O\left(1/\sqrt{N}\right)\right)^2 > (2-\gamma)^2 - O\left(1/\sqrt{N}\right). \tag{H.33}$$

□

H.2 Proofs for deep random features

In this section, we provide the proofs for our results on the deep random features model. We consider a single textual data-point $X = [x_1, x_2, ..., x_N]^\top \in \mathbb{R}^{N\times d}$ that satisfies Assumption 7.1. We consider the deep random features model defined in (7.6), i.e.,

$$\varphi_{\mathsf{DRF}}(X) := \phi(V_L\phi(V_{L-1}(...V_2\phi(V_1 \operatorname{flat} X)...))), \tag{H.34}$$

where $\phi : \mathbb{R} \to \mathbb{R}$ is the non-linearity applied component-wise at each layer, and $V_l \in \mathbb{R}^{D\times D}$ are the random weights at layer l, sampled independently and such that $[V_1]_{i,j} \sim_{\text{i.i.d.}} \mathcal{N}(0, \beta/D)$ and $[V_l]_{i,j} \sim_{\text{i.i.d.}} \mathcal{N}(0, \beta/k)$ for $l > 1$. We set β according to He's (or Kaiming) initialization He et al. [2015], i.e.,

$$\mathbb{E}_{\rho\sim\mathcal{N}(0,\beta)}\left[\phi^2(\rho)\right] = 1. \tag{H.35}$$

Thus, we require the activation function ϕ to guarantee at least one value β for which the previous equation is respected. We will consider β to be a positive constant dependent only on the activation ϕ.

Let's introduce the shorthands

$$\begin{aligned} &\varphi_0(X) = \text{flat}(X), \\ &\varphi_l(X) = \phi\left(V_l\, \varphi_{l-1}(X)\right), \qquad \text{for } l \in [L]. \end{aligned} \tag{H.36}$$

The outline of this section is the following:

1. In Lemma H.4, we prove that at every layer, the norm of the features $\|\varphi_l(X)\|_2$, with $l > 1$, concentrates to $\sqrt{k}$. This is the step where He's initialization is necessary.
2. In Lemma H.5, we show that $\|\varphi_l(X^i(\Delta)) - \varphi_l(X)\|_2$ can be upper bounded by a term that grows exponentially with the depth of the layer l.
3. In Theorem 7.4, we upper bound $\mathcal{S}_{\text{DRF}}$, concluding the argument.

Lemma H.4. *Let $\varphi_l(X)$ be defined in* (H.36)*, and let ϕ be a Lipschitz function such that* (7.7) *admits at least one solution β. Let $X \in \mathbb{R}^{N\times d}$ be a generic input sample such that Assumption 7.1 holds, and let $k = \Theta(D)$. Then, for every $l > 0$, we have*

$$\left|\|\varphi_l(X)\|_2 - \sqrt{k}\right| \le e^{Cl}\log D, \tag{H.37}$$

with probability at least $1 - 2L\exp(-c\log^2 D)$ over $\{V_l\}_{l=1}^L$.

Proof. By Assumption 7.1 the statement trivially holds for $l = 0$, if we replace $\sqrt{k}$ with $\sqrt{d}$. Let's consider this the base case of an induction argument to prove the statement for $l \in [L]$.

Thus, using the notation $k_0 = D$ and $k_l = k$ for $l \in [L]$, the inductive hypothesis becomes

$$\left|\|\varphi_{l-1}(X)\|_2 - \sqrt{k_{l-1}}\right| \le e^{C(l-1)}\log D, \tag{H.38}$$

with probability at least $1 - 2(l-1)\exp(-c\log^2 D)$ on $\{V_m\}_{m=1}^{l-1}$. Let's condition on this high probability event and on $\|V_l\|_{\text{op}} \le C_1$ until the end of the proof. By Theorem 4.4.5 of Vershynin [2018] and since $k = \Theta(D)$, there exists C_1 large enough and independent from l, such that this holds with probability at least $1 - 2\exp(-c_1 k)$ over V_l. We therefore aim to prove the thesis for

$$e^C := \max\left(1, MC_1 + 1\right), \tag{H.39}$$

where we denote by M the Lipschitz constant of ϕ. (H.39) explicitely shows that C is a natural constant dependent only on the activation function ϕ.

We have

$$\varphi_l(X) = \phi(V_l\,\varphi_{l-1}(X)) = \phi\left(V_l\frac{\sqrt{k_{l-1}}\,\varphi_{l-1}(X)}{\|\varphi_{l-1}(X)\|_2} + V_l\left(\varphi_{l-1}(X) - \frac{\sqrt{k_{l-1}}\,\varphi_{l-1}(X)}{\|\varphi_{l-1}(X)\|_2}\right)\right), \tag{H.40}$$

which gives

$$\left\|\varphi_l(X) - \phi\left(V_l \frac{\sqrt{k_{l-1}}\,\varphi_{l-1}(X)}{\|\varphi_{l-1}(X)\|_2}\right)\right\|_2 \leq MC_1 \left|\|\varphi_{l-1}(X)\|_2 - \sqrt{k_{l-1}}\right| \leq MC_1 e^{C(l-1)} \log D, \tag{H.41}$$

where the first step is true since ϕ is M-Lipschitz, and the last step is a direct consequence of the inductive hypothesis (H.38).

Let's now consider the second term in the left hand side of (H.41). Let's define the shorthand $\rho = V_l \frac{\sqrt{k_{l-1}}\,\varphi_{l-1}(X)}{\|\varphi_{l-1}(X)\|_2} \in \mathbb{R}^D$. In the probability space of V_l, ρ is distributed as a Gaussian random vector, such that all its entries ρ_i are i.i.d. Gaussian with variance β. Thus, we have

$$\mathbb{E}_{V_l}\left[\left\|\phi\left(V_l \frac{\sqrt{k_{l-1}}\,\varphi_{l-1}(X)}{\|\varphi_{l-1}(X)\|_2}\right)\right\|_2^2\right] = \mathbb{E}_\rho\left[\|\phi(\rho)\|_2^2\right] = k_l \mathbb{E}_{\rho_1}\left[\phi^2(\rho_1)\right] = k, \tag{H.42}$$

where the last step is a consequence of (H.35), and of $k_l = k$ for $l \in [L]$.

As the ρ_i's are independent and ϕ is Lipschitz, the random variables $(\phi^2(\rho_i) - 1)$ are independent, mean-0, and sub-exponential, such that $\|\phi^2(\rho_i) - 1\|_{\psi_1} \leq C_2$. Thus, by Bernstein inequality (cf. Theorem 2.8.1. in Vershynin [2018]), we have

$$\mathbb{P}_\rho\left(\left|\sum_{i=1}^k \left(\phi^2(\rho_i) - 1\right)\right| \geq \sqrt{k}\log k\right) \leq 2\exp(-c_2 \log^2 k), \tag{H.43}$$

which gives

$$\left|\left\|\phi\left(V_l \frac{\sqrt{k}\,\varphi_{l-1}(X)}{\|\varphi_{l-1}(X)\|_2}\right)\right\|_2^2 - k\right| \leq \sqrt{k}\log k, \tag{H.44}$$

with probability at least $1 - 2\exp(-c_2 \log^2 k)$ over V_l. We will condition on such high probability event until the end of the proof.

Thus, we have

$$\left\|\phi\left(V_l \frac{\sqrt{k}\,\varphi_{l-1}(X)}{\|\varphi_{l-1}(X)\|_2}\right)\right\|_2 \leq \sqrt{k + \sqrt{k}\log k} = \sqrt{k}\sqrt{1 + \frac{\log k}{\sqrt{k}}} \leq \sqrt{k}\left(1 + \frac{\log k}{\sqrt{k}}\right) = \sqrt{k} + \log k, \tag{H.45}$$

and

$$\left\|\phi\left(V_l \frac{\sqrt{k}\,\varphi_{l-1}(X)}{\|\varphi_{l-1}(X)\|_2}\right)\right\|_2 \geq \sqrt{k - \sqrt{k}\log k} = \sqrt{k}\sqrt{1 - \frac{\log k}{\sqrt{k}}} \geq \sqrt{k}\left(1 - \frac{\log k}{\sqrt{k}}\right) = \sqrt{k} - \log k. \tag{H.46}$$

Putting together the last two equations gives

$$\left|\left\|\phi\left(V_l \frac{\sqrt{k}\,\varphi_{l-1}(X)}{\|\varphi_{l-1}(X)\|_2}\right)\right\|_2 - \sqrt{k}\right| \leq \log k. \tag{H.47}$$

Applying (H.41), (H.47) and the triangle inequality gives

$$\left|\|\varphi_l(X)\|_2 - \sqrt{k}\right| \leq MC_1 e^{C(l-1)} \log k + \log k \leq \left(MC_1 e^{C(l-1)} + e^{C(l-1)}\right)\log k \leq e^{Cl} \log k, \tag{H.48}$$

where the second and the third step are both consequences of (H.39). This inequality, performing a union bound on the high probability events we considered so far, holds with probability at least $1 - 2(l-1)\exp(-c\log^2 k) - 2\exp(-c_1 k) - 2\exp(-c_2 \log^2 k) \geq 1 - 2l\exp(-c\log^2 k)$ over $\{V_m\}_{m=1}^l$, as soon as we consider $c = \min(c_1, c_2)$. □

Lemma H.5. *Let $\varphi_l(X)$ be defined in* (H.36)*, let ϕ be a Lipschitz function and $X \in \mathbb{R}^{N\times d}$ a generic input sample such that Assumption 7.1 holds. Assume $k = \Theta(D)$. Then, for every $l \geq 0$, for every $i \in [N]$ and for any $\Delta \in \mathbb{R}^d$ such that $\|\Delta\|_2 \leq \sqrt{d}$, we have*

$$\left\|\varphi_l(X^i(\Delta)) - \varphi_l(X)\right\|_2 \leq \sqrt{d}e^{Cl}, \tag{H.49}$$

with probability at least $1 - 2L\exp(-ck)$ over $\{V_l\}_{l=1}^L$.

Proof. Let's prove the statement by induction over l. The base case $l = 0$ is a direct consequence of $\|\mathrm{flat}(X^i(\Delta)) - \mathrm{flat}(X)\|_2 = \|\Delta\|_2 \leq \sqrt{d}$, which makes the thesis true for any $C \geq 0$.

In the inductive step, the inductive hypothesis becomes

$$\left\|\varphi_{l-1}(X^i(\Delta)) - \varphi_{l-1}(X)\right\|_2 \leq \sqrt{d}e^{C(l-1)}, \tag{H.50}$$

with probability at least $1 - 2(l-1)\exp(-ck)$ on $\{V_m\}_{m=1}^{l-1}$. Let's condition on this high probability event and on $\|V_l\|_{\mathsf{op}} \leq C_1$ until the end of the proof. By Theorem 4.4.5 of Vershynin [2018] and since $k = \Theta(D)$, there exists C_1 large enough and independent from l, such that this holds with probability at least $1 - 2\exp(-c_1 k)$ over V_l.

We aim to prove the thesis for $e^C = MC_1$, and $c = c_1$, where M is the Lipschitz constant of ϕ. We have

$$\begin{aligned}\left\|\varphi_l(X^i(\Delta)) - \varphi_l(X)\right\|_2 &= \left\|\phi(V_l\varphi_{l-1}(X^i(\Delta))) - \phi(V_l\varphi_{l-1}(X))\right\|_2 \\ &\leq MC_1\left\|\varphi_{l-1}(X^i(\Delta)) - \varphi_{l-1}(X)\right\|_2 \\ &\leq MC_1\sqrt{d}e^{C(l-1)} \\ &\leq \sqrt{d}e^{Cl},\end{aligned} \tag{H.51}$$

with probability at least $1-2(l-1)\exp(-ck)-2\exp(-c_1k) = 1-2l\exp(-ck)$ over $\{V_m\}_{m=1}^l$, which gives the thesis. □

Theorem 7.3. Let $\varphi_{\mathsf{DRF}}(X)$ be the deep random feature map defined in (7.6), and let ϕ be a Lipschitz function, such that (7.7) admits at least one solution β. Let $X \in \mathbb{R}^{N\times d}$ be a generic input sample s.t. Assumption 7.1 holds, and assume $k = \Theta(D)$, $L = o(\log k)$. Let $\mathcal{S}_{\mathsf{DRF}}(X)$ be the the word sensitivity defined in (7.4). Then, we have

$$\mathcal{S}_{\mathsf{DRF}} = O\left(\frac{e^{CL}}{\sqrt{N}}\right), \tag{H.52}$$

with probability at least $1 - \exp(-c\log^2 k)$ over $\{V_m\}_{m=1}^l$.

Proof. By Lemma H.4 (for the case $l = L$) and since $L = o(\log k)$, we have

$$\|\varphi_{\mathsf{DRF}}(X)\|_2 = \Theta(\sqrt{k}), \tag{H.53}$$

with probability at least $1 - 2L\exp(-c_1\log^2 k) \geq 1 - \exp(-c_2\log^2 k)$ over $\{V_l\}_{l=1}^L$.

By Lemma H.5 (for the case $l = L$), we have that, for every $i \in [N]$, we have

$$\sup_{\|\Delta\|_2\leq\sqrt{d}}\left\|\varphi_{\mathsf{DRF}}(X^i(\Delta)) - \varphi_{\mathsf{DRF}}(X)\right\|_2 \leq \sqrt{d}e^{CL}, \tag{H.54}$$

with probability at least $1 - 2L\exp(-c_3k) \geq 1 - \exp(-c_4\log^2 k)$ over $\{V_l\}_{l=1}^L$.

Then, putting together (H.53), (H.54), and (7.4), and recalling that $k = \Theta(D)$, we get the thesis. □

H.3 Proofs for random attention features

In this section, we provide the proofs for our results on the random attention features model.

We consider a single textual data-point $X = [x_1, x_2, ..., x_N]^\top \in \mathbb{R}^{N\times d}$ that satisfies Assumptions 7.1. We assume $d/\log^4 d = \Omega(N)$. We consider the random attention features model defined in (7.3), i.e.,

$$\varphi_{\mathsf{RAF}}(X) = \mathrm{softmax}\left(\frac{XWX^\top}{\sqrt{d}}\right)X, \tag{H.55}$$

where $W_{i,j} \sim_{\text{i.i.d.}} \mathcal{N}(0, 1/d)$ sampled independently from X. We define the argument of the $\mathrm{softmax}$ as $S(X) \in \mathbb{R}^{N\times N}$ given by

$$S(X) = \frac{XWX^\top}{\sqrt{d}}, \tag{H.56}$$

and its evaluation on the perturbed sample as

$$S(X^i(\Delta)) = \frac{X^i(\Delta)WX^i(\Delta)^\top}{\sqrt{d}}. \tag{H.57}$$

The attention scores are therefore defined as the row-wise $\mathrm{softmax}$ of the previous term, i.e.,

$$[s(X)]_{j:} = \mathrm{softmax}\left([S(X)]_{j:}\right), \tag{H.58}$$

which allows us to rewrite (7.3) as

$$\varphi_{\mathsf{RAF}}(X) = s(X)X. \tag{H.59}$$

Through this section, we will always consider this notation and assumptions, which will therefore not be repeated in the statement of the lemmas before our final result Theorem 7.4.

The outline of this section is the following:

1. In Lemma H.6, we prove the existence of a vector $\delta^* \in \mathbb{R}^d$, such that its inner product with the token embeddings x_i's is large for a constant fraction of the words in the context.

2. In Lemma H.7, we show that the result in the previous Lemma can be extended to two different vectors δ_1^* and δ_2^*, that are very different from each other, i.e., $\|\delta_1^* - \delta_2^*\|_2 = \sqrt{d}$. The reason why we would like this statement to hold on two vectors will be clear in Lemma H.10.

3. In Lemma H.8, exploiting the properties of the Gaussian attention features W, we show that there exist two different vectors Δ_1^* and Δ_2^*, that are very different from each other, i.e., $\|\Delta_1^* - \Delta_2^*\|_2 = \Omega(\sqrt{d})$, such that a constant fraction of the entries of the vector $XW\Delta_k^*/\sqrt{d}$, $k \in \{1, 2\}$, is large.

4. In Lemma H.9, we prove that, for the two vectors Δ_1^* and Δ_2^* and an $\Omega(N)$ number of rows $j \in [N]$, the attention scores $[s(X^i(\Delta_k^*))]_{j:}^\top$ are well approximated by the canonical basis vector e_i for $k \in \{1, 2\}$. This intuitively means that a constant fraction of tokens x_j puts all their attention towards the i-th modified token $x_i + \Delta_k^*$.

5. In Lemma H.10, using an argument by contradiction, we prove that at least one between Δ_1^* and Δ_2^* is such that $\|\varphi_{\mathsf{RAF}}(X) - \varphi_{\mathsf{RAF}}(X^i(\Delta^*))\|_F = \Omega(\sqrt{dn})$.

6. In Theorem 7.4, we upper bound $\|\varphi_{\mathsf{RAF}}(X)\|_F$, concluding the argument.

Lemma H.6. *There exists a vector $\delta^* \in \mathbb{R}^d$, such that*

$$\|\delta^*\|_2 \leq \sqrt{d}, \tag{H.60}$$

and

$$\left(x_i^\top \delta^*\right)^2 = \Omega\left(\frac{d^2}{N}\right), \tag{H.61}$$

for at least $\Omega(N)$ indices $i \in [N]$.

Proof. Let $X^\top = UDV^\top$ the singular value decomposition of $X^\top$, where $U \in \mathbb{R}^{d\times d}$, $D \in \mathbb{R}^{d\times N}$, and $V \in \mathbb{R}^{N\times N}$, and let $U_r \in \mathbb{R}^{d\times r}$ be the matrix obtained keeping only the first $r = \operatorname{rank}(X)$ columns of U. Let δ_ε be defined as follows

$$\delta_\varepsilon := \sqrt{\frac{d}{N}} U_r \varepsilon, \tag{H.62}$$

where we consider $\varepsilon \in \mathbb{R}^r$ uniformly distributed on the sphere $\sqrt{r}\,\mathbb{S}^{r-1}$ of radius $\sqrt{r}$. By definition we have $U_r^\top U_r = I$, which implies

$$\|U_r\varepsilon\|_2 = \|\varepsilon\|_2 = \sqrt{r}. \tag{H.63}$$

As $r \leq N$, we have that

$$\|\delta_\varepsilon\|_2 \leq \sqrt{d}, \tag{H.64}$$

for every ε. Let's call Z_i^ε the random variable $x_i^\top \delta_\varepsilon$.

By contradiction, let's assume the thesis to be false. Then, in particular, by (H.64), there is no ε such that δ_ε respects the second part of the thesis. This implies that, for every ε, there are at least $\lceil N/2 \rceil$ values of i such that $(Z_i^\varepsilon)^2 < 0.01\, d^2/N$. Let's define the indicators

$$\chi_i = \begin{cases} 1, & \text{if } (Z_i^\varepsilon)^2 < 0.01\, d^2/N, \\ 0, & \text{if } (Z_i^\varepsilon)^2 \geq 0.01\, d^2/N. \end{cases} \tag{H.65}$$

Thus, by contradiction, we have

$$\sum \chi_i \geq \left\lceil \frac{N}{2} \right\rceil \geq \frac{N}{2}, \tag{H.66}$$

for every ε. Thus, by the probabilistic method,

$$\frac{N}{2} \leq \mathbb{E}_\varepsilon\left[\sum \chi_i\right] = \sum \mathbb{E}_\varepsilon\left[\chi_i\right] = N\, \mathbb{E}_\varepsilon\left[\chi_1\right] = N\, \mathbb{P}_\varepsilon\left((Z_1^\varepsilon)^2 \leq 0.01\, d^2/N\right), \tag{H.67}$$

where the second equality is true as the Z_i^ε are identically distributed, and the last step comes from the definition of the indicators in (H.65). This implies, for every i,

$$\mathbb{P}_\varepsilon\left((Z_i^\varepsilon)^2 \leq 0.01\, d^2/N\right) \geq 0.5. \tag{H.68}$$

Let $\rho \in \mathbb{R}^r$ be a standard Gaussian vector and define $\rho_r = \|\rho\|_2 / \sqrt{r}$. We have that $\mathbb{E}[\rho_r^2] = 1$ and $\mathrm{Var}(\rho_r^2) = 1/r$. Thus, for every r, by Chebyshev's inequality, we can write

$$\mathbb{P}_{\rho_r}\left(\rho_r^2 \leq 3\right) \geq 0.75 \geq 0.5. \tag{H.69}$$

By rotational invariance of the Gaussian measure, we have that $\rho_r \varepsilon$ is distributed as a standard Gaussian vector. In particular, we have

$$\rho_r Z_i^\varepsilon = \rho_r x_i^\top \delta_\varepsilon = \left(\sqrt{\frac{d}{N}} U_r^\top x_i\right)^\top (\rho_r \varepsilon)\,. \tag{H.70}$$

As $U_r^\top x_i$ is a fixed vector independent from ε and ρ_r, we have that $\rho_r Z_i^\varepsilon$ is a Gaussian random variable (in the probability space of ε and ρ_r), with variance $d \left\|U_r^\top x_i\right\|_2^2 / N = d^2/N$, as $x_i \in \mathrm{Span}\{\mathrm{rows}(U_r^\top)\}$ by construction. Thus, $\sqrt{N}\rho_r Z_i^\varepsilon / d$ is distributed as a standard Gaussian random variable g_i.

Therefore, putting together (H.68) and (H.69) gives

$$\mathbb{P}_{g_i}\left(g_i^2 \leq 0.03\right) \geq 0.25. \tag{H.71}$$

However, we can upper-bound the left hand side of the previous equation exploiting the closed form of the pdf of a standard Gaussian random variable. This gives

$$\mathbb{P}_{g_i}\left(|g_i| \leq \sqrt{0.03}\right) < \frac{1}{\sqrt{2\pi}} \cdot 2 \cdot \sqrt{0.03} < 0.14 < 0.25, \tag{H.72}$$

which leads to the desired contradiction. □

Lemma H.7. *There exist two vectors $\delta_1^* \in \mathbb{R}^d$ and $\delta_2^* \in \mathbb{R}^d$ such that*

$$\|\delta_1^*\|_2 \leq \sqrt{d}, \qquad \|\delta_2^*\|_2 \leq \sqrt{d}, \qquad \|\delta_1^* - \delta_2^*\|_2 = \sqrt{d}, \tag{H.73}$$

and $x_i^\top \delta_1^* > 0$, $x_i^\top \delta_2^* > 0$, and

$$x_i^\top \delta_1^* = \Omega\left(\sqrt{d}\log^2 d\right), \tag{H.74}$$

$$x_i^\top \delta_2^* = \Omega\left(\sqrt{d}\log^2 d\right), \tag{H.75}$$

for at least $\Omega(N)$ indices $i \in [N]$.

Proof. By Lemma H.6, we have that there exists a vector δ^*, with $\|\delta^*\|_2 \leq \sqrt{d}$ such that

$$\left(x_i^\top \delta^*\right)^2 = \Omega\left(\frac{d^2}{N}\right), \tag{H.76}$$

for at least $\Omega(N)$ indices $i \in [N]$. This also implies (up to considering $-\delta^*$ instead of δ^*) that

$$x_i^\top \delta^* = \Omega\left(\frac{d}{\sqrt{N}}\right) = \Omega\left(\sqrt{d}\log^2 d\right), \tag{H.77}$$

where the last step is justified by $d/\log^4 d = \Omega(N)$.

Let's now define

$$\delta_1^* = \frac{\delta^*}{2} + \frac{v}{2}, \tag{H.78}$$

and

$$\delta_2^* = \frac{\delta^*}{2} - \frac{v}{2}, \tag{H.79}$$

where $v \in \mathbb{R}^d$ is a generic fixed vector such that $\|v\|_2 = \sqrt{d}$, and $\|Xv\|_2 = 0$, i.e. $v \in \mathrm{Span}\{\mathrm{rows}(X)\}^\perp$. This is again possible because of $d/\log^4 d = \Omega(N)$. As $\|\delta^*\|_2 \leq \sqrt{d}$ by Lemma H.6, the first part of the thesis follows from a straightforward application of the triangle inequality.

Since $v^\top x_i = 0$ for every $1 \leq i \leq N$, we also have

$$x_i^\top \delta_1^* = x_i^\top \delta_2^* = x_i^\top \delta^*/2, \tag{H.80}$$

which readily gives the desired result. □

Lemma H.8. *Let's define $\sigma(\Delta) \in \mathbb{R}^N$ as*

$$\sigma(\Delta) := \frac{XW\Delta}{\sqrt{d}}, \tag{H.81}$$

where $\Delta \in \mathbb{R}^d$. Then, with probability at least $1 - \exp(-c\log^2 d)$ over W, there exist Δ_1^* and Δ_2^* such that,

$$\|\Delta_1^*\|_2 \leq \sqrt{d}, \qquad \|\Delta_2^*\|_2 \leq \sqrt{d}, \qquad \|\Delta_1^* - \Delta_2^*\|_2 = \Omega(\sqrt{d}), \tag{H.82}$$

and

$$\sigma(\Delta_1^*)_i = \Omega\left(\log^2 d\right). \tag{H.83}$$

$$\sigma(\Delta_2^*)_i = \Omega\left(\log^2 d\right). \tag{H.84}$$

for at least $\Omega(N)$ indices $i \in [N]$.

Proof. Let's set

$$\Delta = CW^\top \delta, \tag{H.85}$$

where $\delta \in \mathbb{R}^d$ is independent from W, and C is an absolute constant that will be fixed later in the proof. For every i, we can write

$$\delta = \frac{\delta^\top x_i}{\|x_i\|_2^2} x_i + \delta_i^\perp, \tag{H.86}$$

with $\delta_i^\perp$ being orthogonal with x_i by construction. Let's for now suppose $\left\|\delta_i^\perp\right\|_2 \neq 0$. Let $a(\delta)_i := \delta^\top x_i$. We have

$$\begin{aligned}
\sigma(\Delta)_i &\geq C\frac{a(\delta)_i}{\sqrt{d}\,\|x_i\|_2^2}\left\|W^\top x_i\right\|_2^2 - C\left|\frac{x_i^\top WW^\top \delta_i^\perp}{\sqrt{d}}\right| \\
&= C\frac{a(\delta)_i}{\sqrt{d}\,\|x_i\|_2^2}\left\|W^\top x_i\right\|_2^2 - C\frac{\left\|\delta_i^\perp\right\|_2}{d}\left|x_i^\top WW^\top \frac{\sqrt{d}\,\delta_i^\perp}{\left\|\delta_i^\perp\right\|_2}\right|.
\end{aligned} \tag{H.87}$$

In the probability space of W, as the entries of W are i.i.d. and Gaussian $\mathcal{N}(0, 1/d)$, we have that $W^\top x_i$ and $W^\top \frac{\sqrt{d}\,\delta_i^\perp}{\left\|\delta_i^\perp\right\|_2}$ are two independent standard Gaussian vectors, namely, ρ_1 and ρ_2. This implies

$$\left\|W^\top x_i\right\|_2^2 = \Omega(d), \qquad \left| x_i^\top W W^\top \frac{\sqrt{d}\,\delta_i^\perp}{\left\|\delta_i^\perp\right\|_2} \right| =: \left|\rho_1^\top \rho_2\right| = O\left(\sqrt{d}\log d\right), \tag{H.88}$$

with probability at least $1 - \exp(-c_1 \log^2 d)$, over the probability space of W. The first statement holds because of Theorem 3.1.1. of Vershynin [2018], and the second one because, in the probability space of ρ_1, $\rho_1^\top \rho_2$ is a Gaussian random variable with variance $\|\rho_2\|_2^2$, which in turn is $O(d)$ with probability at least $1 - \exp(-c_2 d)$ over ρ_2. Using $\|x_i\|_2^2 = d$, we get, for two positive absolute constants C_1 and C_2,

$$\sigma(\Delta)_i \geq C_1 \frac{a(\delta)_i}{\sqrt{d}} - C_2 \frac{\left\|\delta_i^\perp\right\|_2 \log d}{\sqrt{d}} \geq C_1 \frac{a(\delta)_i}{\sqrt{d}} - C_2 \frac{\|\delta\|_2 \log d}{\sqrt{d}}. \tag{H.89}$$

Notice that the previous equation would still hold even in the case $\left\|\delta_i^\perp\right\|_2 = 0$, which is therefore a case now included in our derivation.

By Lemma H.7, we can choose two vectors δ_1^* and δ_2^*, independently from W such that, for $k \in \{1, 2\}$, $\|\delta_k^*\|_2 \leq \sqrt{d}$ and $a(\delta_k^*)_i = \Omega(\sqrt{d}\log^2 d)$ for $\Omega(N)$ indices $i \in [N]$. This gives, for another absolute positive constant C_3,

$$\sigma(\Delta_1^*)_i \geq C_3 \log^2 d - C_2 \log d = \Omega(\log^2 d), \tag{H.90}$$

$$\sigma(\Delta_2^*)_i \geq C_3 \log^2 d - C_2 \log d = \Omega(\log^2 d), \tag{H.91}$$

for $\Omega(N)$ indices $i \in [N]$.

Let's now verify that $\|\Delta_1^* - \Delta_2^*\|_2 = C\left\|W^\top(\delta_1^* - \delta_2^*)\right\|_2 = \Omega(\sqrt{d})$. As $\delta_1^* - \delta_2^*$ is independent on W, and $\|\delta_1^* - \delta_2^*\|_2 = \sqrt{d}$, we have that $W^\top(\delta_1^* - \delta_2^*)$ is a standard Gaussian random vector, in the probability space of W. Thus, by Theorem 3.3.1 of Vershynin [2018], we have

$$\|\Delta_1^* - \Delta_2^*\|_2 = C\left\|W^\top(\delta_1^* - \delta_2^*)\right\|_2 = \Omega(\sqrt{d}), \tag{H.92}$$

with probability at least $1 - \exp(-c_3 d)$, on the probability space of W.

We are left to verify that, for $k \in \{1, 2\}$, $\|\Delta_k^*\|_2 = C\left\|W^\top \delta_k^*\right\|_2 \leq \sqrt{d}$. This is readily implied by Theorem 4.4.5 of Vershynin [2018], which gives $\|W\|_{\mathsf{op}} = O(1)$ with probability at least $1 - \exp(-c_4 d)$. Taking the intersection between this high probability event and the ones in (H.88) and (H.92), and setting C small enough to have $C\|W\|_{\mathsf{op}} \leq 1$, we get the thesis. □

Lemma H.9. *For every $i \in [N]$, with probability at least $1 - \exp(-c\log^2 d)$ over W, there exist Δ_1^* and Δ_2^* such that,*

$$\|\Delta_1^*\|_2 \leq \sqrt{d}, \qquad \|\Delta_2^*\|_2 \leq \sqrt{d}, \qquad \|\Delta_1^* - \Delta_2^*\|_2 = \Omega(\sqrt{d}), \tag{H.93}$$

and

$$\left\|[s(X^i(\Delta_1^*))]_{j:}^\top - e_i\right\|_2 = o\left(\frac{1}{\sqrt{d}}\right), \tag{H.94}$$

$$\left\|[s(X^i(\Delta_2^*))]_{j:}^\top - e_i\right\|_2 = o\left(\frac{1}{\sqrt{d}}\right), \tag{H.95}$$

for at least $\Omega(N)$ indices $j \in [N]$, where e_i represents the i-th element of the canonical basis in $\mathbb{R}^N$.

Proof. As we can write $X^i(\Delta) = X + e_i\Delta^\top$, we can rewrite (H.57) as

$$\sqrt{d}\, S(X^i(\Delta)) = X^i(\Delta)WX^i(\Delta)^\top = XWX^\top + XW\Delta e_i^\top + e_i\Delta^\top WX + e_i\Delta^\top W\Delta e_i^\top. \tag{H.96}$$

Let's look at the j-th row of $S(X^i(\Delta))$, with $j \neq i$. We have

$$[S(X^i(\Delta))]_{j:} = \frac{x_j^\top WX^\top}{\sqrt{d}} + \frac{x_j^\top W\Delta e_i^\top}{\sqrt{d}}. \tag{H.97}$$

With probability at least $1 - \exp(-c_1 \log^2 d)$ over W, we have that all the N entries of the vector $x_j^\top WX^\top/\sqrt{d}$ are smaller than $\log d$, as they are all standard Gaussian random variables. By Lemma H.8, we have that there exist Δ_1^* and Δ_2^* such that $\|\Delta_1^*\|_2 \leq \sqrt{d}$, $\|\Delta_2^*\|_2 \leq \sqrt{d}$, $\|\Delta_1^* - \Delta_2^*\|_2 = \Omega(\sqrt{d})$ and

$$\frac{x_j^\top W\Delta_k^*}{\sqrt{d}} = \Omega(\log^2 d), \qquad k \in \{1, 2\}, \tag{H.98}$$

for $\Omega(N)$ indices $j \in [N]$, with probability at least $1 - \exp(-c_2 \log^2 d)$ over W. Let's consider this set of indices until the end of the proof, where we have that, for $k \in \{1, 2\}$, $[S(X^i(\Delta_k^*))]_{ji} = \Omega(\log^2 d)$.

After applying the $\mathrm{softmax}$ function to the row in (H.97) as indicated in (H.58), we get, for every index $k \in [N]$ and $k \neq i$

$$\begin{aligned}
[s(X^i(\Delta_1^*))]_{jk} &= \mathrm{softmax}\left([S(X^i(\Delta_1^*))]_{j:}\right)_k \\
&= \frac{\exp\left([S(X^i(\Delta_1^*))]_{jk}\right)}{\sum_l \exp\left([S(X^i(\Delta_1^*))]_{jl}\right)} \\
&\leq \frac{d}{\exp\left([S(X^i(\Delta_1^*))]_{ji}\right)} \\
&\leq \frac{d}{Cd^{\log d}} = o\left(\frac{1}{d}\right),
\end{aligned} \tag{H.99}$$

and

$$\begin{aligned}
[s(X^i(\Delta_1^*))]_{ji} &= \mathrm{softmax}\left([S(X^i(\Delta_1^*))]_{j:}\right)_i \\
&= 1 - \frac{\sum_{l\neq i} \exp\left([S(X^i(\Delta_1^*))]_{jl}\right)}{\sum_l \exp\left([S(X^i(\Delta_1^*))]_{jl}\right)} \\
&\geq 1 - \frac{Nd}{\exp\left([S(X^i(\Delta_1^*))]_{ji}\right)} \\
&\geq 1 - \frac{Nd}{Cd^{\log d}} = 1 - o\left(\frac{1}{d}\right),
\end{aligned} \tag{H.100}$$

where the last step is justified by the fact that $d/\log^4 d = \Omega(N)$.

Thus, for the same reason, we have

$$\left\|[s(X^i(\Delta_1^*))]_{j:}^\top - e_i\right\|_2^2 = \sum_{k\neq i}[s(X^i(\Delta_1^*))]_{jk}^2 + \left(1 - [s(X^i(\Delta_1^*))]_{ji}\right)^2 = No\left(\frac{1}{d^2}\right) = o\left(\frac{1}{d}\right). \tag{H.101}$$

As (H.101) can be written also with respect to Δ_2^*, the thesis readily follows. □

Lemma H.10. *For every $i \in [N]$, with probability at least $1 - \exp(-c \log^2 d)$ over W, there exist Δ^* such that $\|\Delta^*\|_2 \leq \sqrt{d}$, and*

$$\left\|\varphi_{\mathsf{RAF}}(X) - \varphi_{\mathsf{RAF}}(X^i(\Delta^*))\right\|_F = \Omega(\sqrt{dn}). \tag{H.102}$$

Proof. By (H.59), we have that,

$$\varphi_{\mathsf{RAF}}(X^i(\Delta)) = s(X^i(\Delta))X^i(\Delta), \tag{H.103}$$

By Lemma H.9, we have that there exists Δ_1^*, with $\|\Delta_1^*\|_2 \leq \sqrt{d}$, such that there are $\Omega(N)$ indices j such that

$$\left\|[s(X^i(\Delta_1^*))]_{j:}^\top - e_i\right\|_2 = o\left(\frac{1}{\sqrt{d}}\right), \tag{H.104}$$

with probability at least $1 - \exp(-c_1 \log^2 d)$ over W. Until the end of the proof, we will refer to j as a generic index for which the previous equation holds, and we will condition on the high probability event that the equations in the statement of Lemma H.9 hold.

As we can write

$$\left[\varphi_{\mathsf{RAF}}(X^i(\Delta_1^*))\right]_{j:} = [s(X^i(\Delta_1^*))]_{j:}X^i(\Delta_1^*) = e_i^\top X^i(\Delta_1^*) + \left([s(X^i(\Delta_1^*))]_{j:} - e_i^\top\right) X^i(\Delta_1^*), \tag{H.105}$$

and $e_i^\top X^i(\Delta_1^*) = x_i^\top + \Delta_1^{*\top}$, we have

$$\begin{aligned}\left\|\left[\varphi_{\mathsf{RAF}}(X^i(\Delta_1^*))\right]_{j:} - (x_i + \Delta_1^*)^\top\right\|_2 &\leq \left\|[s(X^i(\Delta_1^*))]_{j:} - e_i^\top\right\|_2 \left\|X^i(\Delta_1^*)\right\|_{\mathsf{op}} \\ &\leq o\left(\frac{1}{\sqrt{d}}\right)(\|X\|_F + \|\Delta_1^*\|_2) \\ &= o\left(\frac{1}{\sqrt{d}}\right) O\left(\sqrt{dn}\right) = o\left(\sqrt{d}\right),\end{aligned} \tag{H.106}$$

where the last step follows from our assumption $d/\log^4 d = \Omega(N)$. By Lemma H.9, there also exists Δ_2^* such that $\|\Delta_2^*\|_2 \leq \sqrt{d}$, $\|\Delta_1^* - \Delta_2^*\|_2 = \Omega(\sqrt{d})$ and (H.106) holds.

Let's now suppose by contradiction that, for some j,

$$\begin{aligned}\left\|[\varphi_{\mathsf{RAF}}(X)]_{j:} - \left[\varphi_{\mathsf{RAF}}(X^i(\Delta_1^*))\right]_{j:}\right\|_2 &= o\left(\sqrt{d}\right), \\ \left\|[\varphi_{\mathsf{RAF}}(X)]_{j:} - \left[\varphi_{\mathsf{RAF}}(X^i(\Delta_2^*))\right]_{j:}\right\|_2 &= o\left(\sqrt{d}\right).\end{aligned} \tag{H.107}$$

Then, we have

$$\begin{aligned}\|\Delta_1^* - \Delta_2^*\|_2 &= \|(x_i + \Delta_1^*) - (x_i + \Delta_2^*)\|_2 \\ &= \| \left((x_i + \Delta_1^*) - \left[\varphi_{\mathsf{RAF}}(X^i(\Delta_1^*))\right]_{j:}\right) + \left(\left[\varphi_{\mathsf{RAF}}(X^i(\Delta_1^*))\right]_{j:} - [\varphi_{\mathsf{RAF}}(X)]_{j:}\right) \\ &\quad - \left((x_i + \Delta_2^*) - \left[\varphi_{\mathsf{RAF}}(X^i(\Delta_2^*))\right]_{j:}\right) - \left(\left[\varphi_{\mathsf{RAF}}(X^i(\Delta_2^*))\right]_{j:} - [\varphi_{\mathsf{RAF}}(X)]_{j:}\right) \| \\ &\leq \left\|(x_i + \Delta_1^*) - \left[\varphi_{\mathsf{RAF}}(X^i(\Delta_1^*))\right]_{j:}\right\|_2 + \left\|\left[\varphi_{\mathsf{RAF}}(X^i(\Delta_1^*))\right]_{j:} - [\varphi_{\mathsf{RAF}}(X)]_{j:}\right\|_2 \\ &\quad + \left\|(x_i + \Delta_2^*) - \left[\varphi_{\mathsf{RAF}}(X^i(\Delta_2^*))\right]_{j:}\right\|_2 + \left\|\left[\varphi_{\mathsf{RAF}}(X^i(\Delta_2^*))\right]_{j:} - [\varphi_{\mathsf{RAF}}(X)]_{j:}\right\|_2 \\ &= o(\sqrt{d}),\end{aligned} \tag{H.108}$$

where the third step holds by triangle inequality, and the last comes from (H.106) and (H.107). As $\|\Delta_1^* - \Delta_2^*\|_2 = \Omega(\sqrt{d})$ by Lemma H.9, we get the desired contradiction, and we have that at least one equation in (H.107) doesn't hold.

Let's therefore denote by Δ^* the vector with $\|\Delta^*\|_2 \leq \sqrt{d}$ such that

$$\left\| [\varphi_{\mathsf{RAF}}(X)]_{j:} - \left[\varphi_{\mathsf{RAF}}(X^i(\Delta^*))\right]_{j:} \right\|_2 = \Omega\left(\sqrt{d}\right) \tag{H.109}$$

holds for $\Omega(N)$ indices j. Due to the previous conditioning, such a vector exists with probability at least $1-\exp(-c_1 \log^2 d)$ over W. Let's denote the set of these indices by J, with $|J| = \Omega(N)$ (where we denote set cardinality by $|\cdot|$).

Thus, we have

$$\left\|\varphi_{\mathsf{RAF}}(X) - \varphi_{\mathsf{RAF}}(X^i(\Delta^*))\right\|_F^2 \geq \sum_{j\in J} \left\| [\varphi_{\mathsf{RAF}}(X)]_{j:} - \left[\varphi_{\mathsf{RAF}}(X^i(\Delta^*))\right]_{j:} \right\|_2^2 = \Omega(Nd), \tag{H.110}$$

which concludes the proof.

□

Theorem 7.4 Let $\varphi_{\mathsf{RAF}}(X)$ be the random attention features map defined in (7.3). Let $X \in \mathbb{R}^{N\times d}$ be a generic input sample s.t. Assumption 7.1 holds, and assume $d/\log^4 d = \Omega(N)$. Let $\mathcal{S}_{\mathsf{RAF}}(X)$ be the the word sensitivity defined in (7.4). Then, we have

$$\mathcal{S}_{\mathsf{RAF}}(X) = \Omega(1), \tag{H.111}$$

with probability at least $1 - \exp(-c\log^2 d)$ over W.

Proof. We have $\varphi_{\mathsf{RAF}}(X) = s(X)X$, which implies

$$[\varphi_{\mathsf{RAF}}(X)]_{j:} = [s(X)]_{j:} X, \tag{H.112}$$

and

$$\left\| [\varphi_{\mathsf{RAF}}(X)]_{j:} \right\|_2 = \left\| \sum_{k=1}^N [s(X)]_{jk} x_k \right\|_2 \leq \sum_{k=1}^N [s(X)]_{jk} \|x_k\|_2 = \sqrt{d}, \tag{H.113}$$

where the second step follows from triangle inequality, and the last step holds as $\sum_{k=1}^N [s(X)]_{jk} = 1$, and $\|x_k\|_2 = \sqrt{d}$ for every k. This readily implies

$$\|\varphi_{\mathsf{RAF}}(X)\|_F \leq \sqrt{Nd}. \tag{H.114}$$

By Lemma H.10, we have that, with probability at least $1 - \exp(-c\log^2 d)$ over W, there exist Δ^* such that $\|\Delta^*\|_2 \leq \sqrt{d}$, and

$$\left\|\varphi_{\mathsf{RAF}}(X) - \varphi_{\mathsf{RAF}}(X^i(\Delta^*))\right\|_F = \Omega(\sqrt{dn}). \tag{H.115}$$

This, together with (H.114), concludes the proof. □

H.4 Assumption 7.5 and adversarial robustness

Assumption 7.5 requires the perturbation Δ to be such that the model $f_{\mathsf{RF}}(\cdot, \theta^*)$ gives a similar output when evaluated on the two new samples X and $X^i(\Delta)$. This assumption is necessary to understand the different behaviour of the RF and the RAF model in our setting, as there in fact exists an adversarial patch Δ such that, for example, $f(X^i(\Delta), \theta_r^*)$ and $f(X, \theta_r^*)$ are very different from each other (e.g., $f_{\mathsf{RF}}(X^i(\Delta), \theta_r^*) = y_\Delta$ while $f_{\mathsf{RF}}(X, \theta_r^*) = y$).

This conclusion derives from the adversarial vulnerability of the RF model, extensively studied in previous work Dohmatob and Bietti [2022], Dohmatob [2022], Bombari et al. [2023]. We remark that this vulnerability depends on the scalings of the problem, i.e., N, d and n. In fact, (using the re-trained solution as example) we can write (assuming $\theta_0 = 0$ for simplicity)

$$\begin{aligned}\left|f_{\mathsf{RF}}(X^i(\Delta), \theta_r^*) - f_{\mathsf{RF}}(X, \theta_r^*)\right| &= \left|\left(\varphi_{\mathsf{RF}}(X^i(\Delta)) - \varphi_{\mathsf{RF}}(X)\right)^\top \Phi_{\mathsf{RF},r}^+ \mathcal{Y}_r\right| \\ &\leq \left\|\varphi_{\mathsf{RF}}(X^i(\Delta)) - \varphi_{\mathsf{RF}}(X)\right\|_2 \left\|\Phi_{\mathsf{RF},r}^+\right\|_{\mathsf{op}} \|Y_r\|_2 .\end{aligned} \tag{H.116}$$

If we bound the three terms on the RHS separately, we get:

- $\|\varphi_{\mathsf{RF}}(X^i(\Delta)) - \varphi_{\mathsf{RF}}(X)\|_2 = O\left(\sqrt{k/N}\right)$ with probability at least $1 - \exp(-cD)$ over V, by Lemma H.1;
- $\left\|\Phi_{\mathsf{RF},r}^+\right\|_{\mathsf{op}} = \lambda_{\min}^{-1/2}(K_{\mathsf{RF},r}) = O\left(\sqrt{1/k}\right)$, with probability at least $1 - \exp\left(-c\log^2 n\right)$ over V and $\mathcal{X}_r$, by Lemma H.2;
- $\|Y_r\|_2 = \sqrt{n+1}$, as we are considering labels in $\{-1, 1\}$.

Thus, we conclude that

$$\left|f_{\mathsf{RF}}(X^i(\Delta), \theta_r^*) - f_{\mathsf{RF}}(X, \theta_r^*)\right| = O\left(\sqrt{\frac{n}{N}}\right), \tag{H.117}$$

with high probability. As a consequence, in the regime where $n = o(N)$, $f_{\mathsf{RF}}(\cdot, \theta_r^*)$ cannot distinguish between the samples X and $X^i(\Delta)$, without the additional need for Assumption 7.5. This result is also shown in our experiments, in the left subplots of Figure 7.4, where the points approach smaller values of γ and consequently higher values of the error as n decreases, becoming comparable with N.

H.5 Further experiments

In Figure 7.4, we compute Δ^* by optimizing with respect to it the following two losses (for fine-tuning and re-training, respectively):

$$\ell_{\theta_f^*}(\Delta) := \left(\frac{\varphi_{\mathsf{RAF}}(X^i(\Delta))^\top \varphi_{\mathsf{RAF}}(X)}{\|\varphi_{\mathsf{RAF}}(X)\|_2^2} + 1\right)^2, \tag{H.118}$$

$$\ell_{\theta_r^*}(\Delta) := \left(\mathcal{F}_{\mathsf{RAF}}(X, X^i(\Delta)) + 1\right)^2, \tag{H.119}$$

subject to the constraint $\|\Delta^*\| \leq \sqrt{d}$. We also report with cross markers the points obtained optimizing the loss $\ell_{\mathsf{Err}}(\Delta) := \mathsf{Err}_{\mathsf{RAF}}(X^i(\Delta), \theta_{f/r}^*)$, showing that this method provides less interesting results, as the error tends to be minimized at the expenses of a large value of γ.

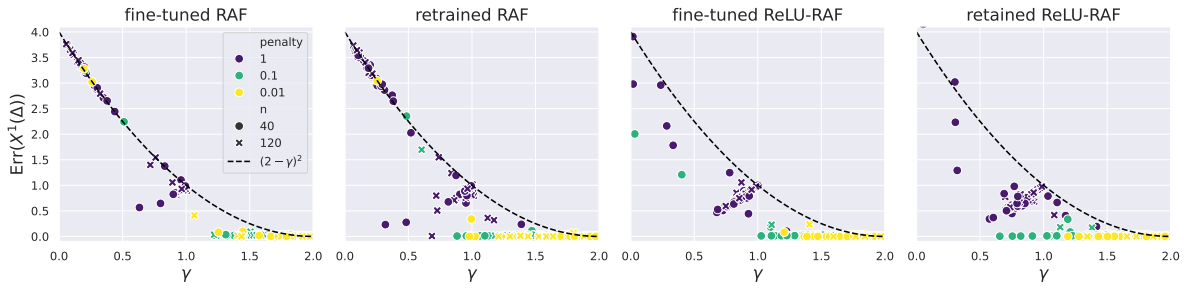


Figure H.1: $\mathrm{Err}_{\varphi}(X^i(\Delta), \theta^*_{f/r})$ for the RAF (two left sub-plots) and ReLU-RAF (two right subplots) maps, as a function of the smallest γ for which Assumption 7.5 is satisfied. Every sub-plot has a fixed context length $N = \{40, 120\}$, embedding dimension $d = 768$ and number of training samples $n = 400$. Every point in the scatter-plots represents an independent simulation where (X, y) and $(\mathcal{X}, \mathcal{Y})$ are the BERT-Base embeddings of a random subset of the imdb dataset (after pre-processing to fulfill Assumption 7.1). For every point, Δ is obtained through constrained gradient descent optimization of $\ell_{\mathsf{Err},p}(\Delta)$, defined in (H.120), for different values of the penalty term $p = \{1, 0.1, 0.01\}$.

An alternative approach is to still optimize with respect to the errors directly, but after introducing a penalty term p on the value of γ. In particular, we consider the following penalized loss

$$\ell_{\mathsf{Err},p}(\Delta) := \mathsf{Err}_{\mathsf{RAF}}(X^i(\Delta), \theta^*_{f/r}) + p\left(f_{\mathsf{RAF}}(X^i(\Delta), \theta^*) - f_{\mathsf{RAF}}(X, \theta^*)\right)^2. \tag{H.120}$$

We perform new experiments with this optimization algorithm, for different values of p, and we report the results in Figure H.1. In this case, compared to the loss $\ell_{\mathsf{Err}}(\Delta)$, we can more easily obtain points that lie below the lower bound. However, it remains difficult to find points where both the error and γ are small.

Another optimization option is to introduce a penalty term to the losses in (H.118) and (H.119):

$$\ell_{\theta^*_f,p}(\Delta) := \left(\frac{\varphi_{\mathsf{RAF}}(X^i(\Delta))^\top \varphi_{\mathsf{RAF}}(X)}{\|\varphi_{\mathsf{RAF}}(X)\|_2^2} + 1\right)^2 + p\left(f_{\mathsf{RAF}}(X^i(\Delta), \theta^*) - f_{\mathsf{RAF}}(X, \theta^*)\right)^2 \tag{H.121}$$

and

$$\ell_{\theta^*_r,p}(\Delta) := \left(\mathcal{F}_{\mathsf{RAF}}(X, X^i(\Delta)) + 1\right)^2 + p\left(f_{\mathsf{RAF}}(X^i(\Delta), \theta^*) - f_{\mathsf{RAF}}(X, \theta^*)\right)^2, \tag{H.122}$$

for the fine-tuning and re-training case respectively. We perform new experiments with this optimization algorithm, for different values of p, and we report the results in Figure H.2. In this case, it is easier to obtain final points that respect Assumption 7.5 with a lower value of γ.

Finally, we consider swapping the losses $\ell_{\theta^*_f}$ and $\ell_{\theta^*_r}$ defined above, i.e., employ the former for re-training and the latter for fine-tuning. Given the heuristic nature of these losses, it is a priori not obvious that they perform their best on the respective setting, as they could be interchangeable. In Figure H.3, we report the results of this investigation. We report in deep-blue the points resulting from the optimization of $\ell_{\theta^*_f}$, and in yellow the points resulting from the optimization of $\ell_{\theta^*_r}$, for both the fine-tuned and re-trained setting. We note that $\ell_{\theta^*_f}$ performs better on the fine-tuned solution, and $\ell_{\theta^*_r}$ better on the re-trained one.

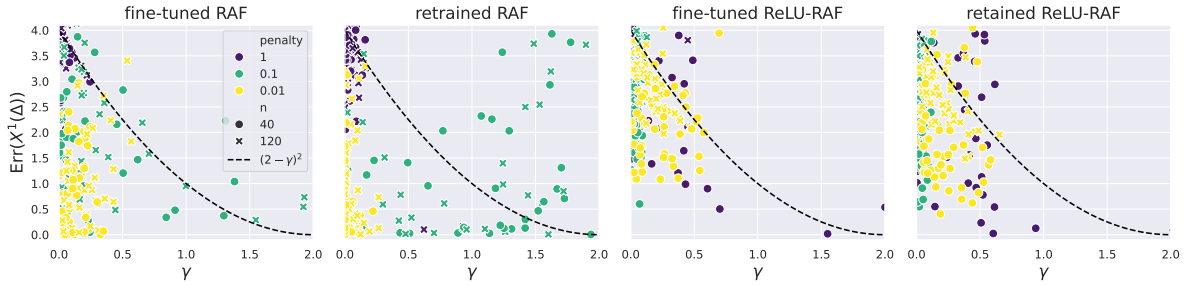


Figure H.2: $\mathrm{Err}_{\varphi}(X^i(\Delta), \theta^*_{f/r})$ for the RAF (two left sub-plots) and ReLU-RAF (two right subplots) maps, as a function of the smallest γ for which Assumption 7.5 is satisfied. For every point, Δ is obtained through constrained gradient descent optimization of $\ell_{\theta^*_r, f}(\Delta)$ in the fine-tuned case, and $\ell_{\theta^*_r, p}(\Delta)$ in the re-trained case, defined in (H.121) and (H.122) respectively, for different values of the penalty term $p = \{1, 0.1, 0.01\}$. The rest of the setup is equivalent to the one described in Figure H.1.

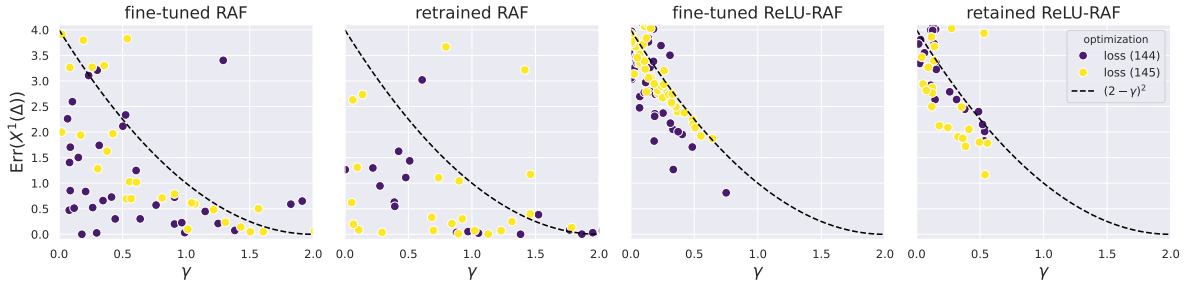


Figure H.3: $\mathrm{Err}_{\varphi}(X^i(\Delta), \theta^*_{f/r})$ for the RAF (two left sub-plots) and ReLU-RAF (two right subplots) maps, as a function of the smallest γ for which Assumption 7.5 is satisfied. We consider fixed embedding dimension $d = 768$, context length $N = 120$, and number of training samples $n = 400$. Every point in the scatter-plots represents an independent simulation where (X, y) and $(\mathcal{X}, \mathcal{Y})$ are the BERT-Base embeddings of a random subset of the imdb dataset (after pre-processing to fulfill Assumption 7.1). For every point, Δ is obtained through constrained gradient descent optimization of either $\ell_{\theta^*_f}$, defined in (H.118), or $\ell_{\theta^*_r}$, defined in (H.119).

APPENDIX

Appendix for Chapter 8

I.1 Further discussion on related work

In Section 8.2, we have discussed how existing bounds are unable to tackle an over-parameterized RF model when $\varepsilon = \Theta(1)$. This requires the characterization of different quantities, *i.e.*, the Lipschitz constant of the loss L, the norm of the baseline solution $\|\theta^*\|_2$, the rank of the matrix M, defined as the projector on the column space of $\mathbb{E}_x[\varphi(x)\varphi(x)^\top]$, and the quantity $\|M\theta^*\|_2$. We provide the desired characterization in the following paragraphs.

Estimate of $\|\theta^*\|_2$. Recall that $\theta^* = \Phi^+ Y$, which readily implies that $\|\theta^*\|_2 \leq \|\Phi^+\|_{\mathsf{op}} \|Y\|_2$. An application of Lemma I.11 gives that, with high probability, $\|\Phi^+\|_{\mathsf{op}} = O\left(p^{-1/2}\right)$. Then, as the labels are bounded, we have $\|\theta^*\|_2 = O\left(\sqrt{n/p}\right)$. This upper bound is tight in different settings, *e.g.*, when the labels contain random noise independent from the data. In this case, an argument similar to the one in Theorem 2 of Bombari et al. [2023], together with Lemma I.11, guarantees that $\|\theta^*\|_2 = \Theta(\|\Phi^+\|_F) = \Theta(\sqrt{n/p})$.

Estimate of L via a bounded set $\mathcal{B}$ with radius $\beta = \Theta(\|\theta^*\|)$. Denoting with (x, y) a generic input-label pair, we can write

$$L = \sup_{\theta \in \mathcal{B}} \left\| \nabla_\theta \ell(\varphi(x)^\top \theta - y) \right\|_2 = \sup_{\theta \in \mathcal{B}} 2 \|\varphi(x)\| \left| \varphi(x)^\top \theta - y \right| = \Theta\left(\beta \|\varphi(x)\|_2^2\right), \tag{I.1}$$

where we use Cauchy-Schwartz inequality in the last step and the fact that $\|\varphi(x)\|_2 \beta \gg |y|$ in our data scaling. Then, noting that $\|\varphi(x)\|_2 = \Theta(\sqrt{p})$ (which holds with high probability) and plugging the value $\beta = \Theta(\|\theta^*\|)$ obtained in the previous paragraph, we obtain $L = \Theta(\sqrt{np})$, and consequently $L \|\theta^*\|_2 = \Theta(n)$, as reported in Section 8.2.

Estimate of L via the set $\mathcal{C}$. Alternatively, one could be interested in an RF model with Lipschitz loss, as the clipped loss defined in (8.6) with C_{clip} set as in (8.13). This choice would guarantee that the Lipschitz constant L of the loss scales as $\sqrt{p}$ (excluding poly-logarithmic factors), which corresponds to the norm of the features $\|\varphi(x)\|_2$. In this setting, we can show that the solution of gradient flow on this loss corresponds to the solution obtained by optimizing the quadratic loss, as the dynamics fully happens in the set $\mathcal{C}$ where the two losses coincide (see Lemma 8.10). Thus, this allows to improve the previous estimate by

a factor $\sqrt{n}$, as we now have $L \left\|\theta^*\right\|_2 = \Theta(\sqrt{p}\sqrt{n/p}) = \Theta(\sqrt{n})$. However, even with this improvement, one cannot obtain meaningful guarantees from Jain and Thakurta [2014] in the over-parameterized regime, as discussed in Section 8.2.

Estimate of $\mathrm{rank}(M)$ and $\left\|M\theta^*\right\|_2$ from Song et al. [2021b]. Let M_D be the projector over the column space of $\Phi^\top\Phi$. Then, under sufficient regularity of the data distribution $\mathcal{P}_X$, it can be shown that M_D spans a subspace of the one spanned by M (this exact argument is used also in the proof of Theorem 3.2 of Song et al. [2021b]). Then, since $M_D\theta^* = \theta^*$ and $\mathrm{rank}(M_D) = n$ from Lemma I.11, we also have that $\left\|M\theta^*\right\|_2 = \left\|\theta^*\right\|_2$ and $\min\left(\mathrm{rank}(M), n\right) = n$. Then, in the over-parameterized RF model, the bound of Song et al. [2021b] corresponds to the one in Jain and Thakurta [2014], and the previous considerations hold.

I.2 Proofs for Section 8.3

We start by introducing the notion of *sensitivity*. Let $\mu : \mathcal{D} \to \mathbb{R}^p$ be an arbitrary, deterministic, p-dimensional function, where $\mathcal{D}$ represents the space of datasets. We define its ℓ_2 sensitivity to be

$$\Delta_2\mu = \sup_{D \text{ adjacent with } D'} \left\|\mu(D) - \mu(D')\right\|_2, \tag{I.2}$$

where D adjacent with D' means that the two datasets D and D' differ by only one training sample. The DP-GD algorithm is obtained through the composition of T independent Gaussian mechanisms (see Dwork and Roth [2014], Theorem A.1), as every iteration includes a (clipped) gradient update and a Gaussian noise injection. In this case, we have that Algorithm 8.1 is a randomized mechanism $\mathcal{A}$ consisting of a sequence of adaptive (Gaussian) mechanisms $\mathcal{M}_1, \ldots, \mathcal{M}_T$ where $\mathcal{M}_t \colon \mathbb{R}^p \times \mathcal{D} \to \mathbb{R}^p$ for all $t \in [T]$. One option is then to use the advanced composition Theorem (see Theorem 3.20 in [Dwork and Roth, 2014]) to compute the privacy guarantees of the algorithm $\mathcal{A}$. Unfortunately, this approach is not fruitful in our case, as it would involve an additional $\log T$ term in our final result in Theorem 8.2, which would make the variance of the noise added at each iteration to diverge, in the limit of $\eta \to 0$, $T \to +\infty$, with $\eta T \to \tau$. Thus, to prove Proposition 8.2, we calculate the privacy guarantees through the moment accountant method, also used by Abadi et al. [2016] to show their Theorem 2 (which we restate in Theorem I.1, for the sake of completeness and notation). This allows for a tighter tracking of the privacy budget, exploiting the independence between the Gaussian mechanisms at every iteration of Algorithm 8.1.

While our approach is conceptually similar to the one used in Theorem 1 of Abadi et al. [2016], in contrast with Abadi et al. [2016] we consider full batch gradient descent and make explicit the dependence on the learning rate η, given our choice of scaling for Algorithm 8.1. We rely on the notion of *privacy loss*, which describes the difference that a randomized mechanism has on two adjacent datasets. More specifically, given two adjacent datasets $D, D' \in \mathcal{D}$, a randomized mechanism $\mathcal{M}_t \colon \mathbb{R}^p \times \mathcal{D} \to \mathbb{R}^p$ and an auxiliary input $\theta_{t-1} \in \mathbb{R}^p$, the *privacy loss* at the output θ takes the form

$$\gamma(\theta; \mathcal{M}_t, \theta_{t-1}, D, D') = \log \frac{p\left(\mathcal{M}_t(\theta_{t-1}, D) = \theta\right)}{p\left(\mathcal{M}_t(\theta_{t-1}, D') = \theta\right)}, \tag{I.3}$$

where the notation $p\left(Z = \theta\right)$ indicates the value of the law of a random variable Z evaluated in θ. As in Abadi et al. [2016], we define the λ-th moment $\alpha_{\mathcal{M}_t}(\lambda; \theta_{t-1}, D, D')$ as the logarithm

of the moment generating function of the privacy loss evaluated in λ, *i.e.*,

$$\alpha_{\mathcal{M}_t}(\lambda; \theta_{t-1}, D, D') = \log \mathbb{E}_{\theta \sim \mathcal{M}_t(\theta_{t-1}, D)} \left[\exp(\lambda \gamma(\theta; \mathcal{M}_t, \theta_{t-1}, D, D')) \right]. \tag{I.4}$$

Note that the expectation is taken with respect to the probability distribution of the parameters given by the mechanism applied to the dataset D. Then, we can define

$$\alpha_{\mathcal{M}_t}(\lambda) = \sup_{\theta_{t-1}, D \text{ adjacent with } D'} \alpha_{\mathcal{M}_t}(\lambda; \theta_{t-1}, D, D'), \tag{I.5}$$

where the supremum is taken over all possible θ_{t-1} and all the adjacent datasets D, D'.

We now re-state Theorem 2 of Abadi et al. [2016] which translates the moment accountant to (ε, δ) privacy guarantees. Next, we bound the sensitivity of each iteration of Algorithm 8.1 and, therefore, the corresponding value of $\alpha_{\mathcal{M}_t}(\lambda)$ (see Lemmas I.2 and I.3). The proof of Proposition 8.2 will then follow.

Theorem I.1 (Theorem 2 of Abadi et al. [2016])**.** *Let $\mathcal{A}$ consists of a sequence of independent adaptive mechanisms $\mathcal{M}_1, \ldots, \mathcal{M}_T$ where $\mathcal{M}_t \colon \mathbb{R}^p \times \mathcal{D} \to \mathbb{R}^p$. Let $\alpha_{\mathcal{M}_t}(\lambda)$ be defined according to* (I.5)*. Then, the following properties hold.*

1. ***Composability.*** *For any λ,*
$$\alpha_{\mathcal{A}}(\lambda) \leq \sum_{t=1}^{T} \alpha_{\mathcal{M}_t}(\lambda). \tag{I.6}$$
2. ***Tail bound.*** *For any $\varepsilon > 0$, $\mathcal{A}$ is (ε, δ)-differentially private for*
$$\delta = \inf_{\lambda} \exp(\alpha_{\mathcal{A}}(\lambda) - \lambda \varepsilon). \tag{I.7}$$

Lemma I.2. *Consider the t-th iteration of Algorithm 8.1 on a dataset D without noise,* i.e.*,*

$$\mu_{\theta_{t-1}, D} = \theta_{t-1} - \frac{\eta}{n} \sum_{(x_i, y_i) \in D} g_{C_{\text{clip}}}(x_i, y_i, \theta_{t-1}). \tag{I.8}$$

Then, we have

$$\Delta_2 \mu = \sup_{\theta_{t-1} \in \mathbb{R}^p, D \text{ adjacent with } D'} \left\| \mu_{\theta_{t-1}, D} - \mu_{\theta_{t-1}, D'} \right\|_2 \leq \frac{2\eta C_{\text{clip}}}{n}. \tag{I.9}$$

Proof. We have

$$\begin{aligned}
\Delta_2 \mu &= \frac{\eta}{n} \sup_{\theta_{t-1} \in \mathbb{R}^p, D \text{ adjacent with } D'} \left\| \sum_{(x_i, y_i) \in D} g_{C_{\text{clip}}}(x_i, y_i, \theta_{t-1}) - \sum_{(x_i, y_i) \in D'} g_{C_{\text{clip}}}(x_i, y_i, \theta_{t-1}) \right\|_2 \\
&= \frac{\eta}{n} \sup_{\theta_{t-1} \in \mathbb{R}^p, (x,y), (x',y') \in (\mathcal{X}, \mathcal{Y})} \left\| g_{C_{\text{clip}}}(x, y, \theta_{t-1}) - g_{C_{\text{clip}}}(x', y', \theta_{t-1}) \right\|_2 \\
&\leq \frac{\eta}{n} \sup_{\theta_{t-1} \in \mathbb{R}^p, (x,y), (x',y') \in (\mathcal{X}, \mathcal{Y})} \left(\left\| g_{C_{\text{clip}}}(x, y, \theta_{t-1}) \right\|_2 + \left\| g_{C_{\text{clip}}}(x', y', \theta_{t-1}) \right\|_2 \right) \\
&\leq \frac{2\eta C_{\text{clip}}}{n},
\end{aligned} \tag{I.10}$$

where the last inequality comes from $\left\| g_{C_{\text{clip}}}(x, y, \theta_{t-1}) \right\|_2 \leq C_{\text{clip}}$ for any $(x, y) \in (\mathcal{X}, \mathcal{Y})$ and any $\theta_{t-1} \in \mathbb{R}^p$. □

Lemma I.3. *Let $\mathcal{M}_t$ be the randomized mechanism induced by the t-th iteration of Algorithm 8.1. Then, for every $t \in [T]$ we have*

$$\alpha_{\mathcal{M}_t}(\lambda) = \sup_{\theta_{t-1}, D \text{ adjacent with } D'} \alpha_{\mathcal{M}_t}(\lambda; \theta_{t-1}, D, D') \leq \frac{\eta}{2\sigma^2}\left(\lambda + \lambda^2\right). \tag{I.11}$$

Proof. As $\mathcal{M}_t$ is a Gaussian mechanism with parameter $\varsigma = \sqrt{\eta}\,\frac{2C_{\text{clip}}}{n}\sigma$, we have

$$p\left(\mathcal{M}_t(\theta_{t-1}, D) = \theta\right) = \frac{1}{\sqrt{(2\pi\varsigma^2)^p}} \exp\left(-\frac{\left\|\theta - \mu_{\theta_{t-1},D}\right\|_2^2}{2\varsigma^2}\right), \tag{I.12}$$

where $\mu_{\theta_{t-1},D}$ represents the update without noise

$$\mu_{\theta_{t-1},D} = \theta_{t-1} - \frac{\eta}{n} \sum_{(x_i,y_i)\in D} g_{C_{\text{clip}}}(x_i, y_i, \theta_{t-1}). \tag{I.13}$$

Equivalently, (I.12) also holds for $\mathbb{P}\left(\mathcal{M}_t(\theta_{t-1}, D') = \theta\right)$, with $\mu_{\theta_{t-1},D'} = \theta_{t-1} - \eta g_{\theta_{t-1},D'}$. Then, we have that the privacy loss defined in (I.3) takes the form

$$\begin{aligned}
&\gamma(\theta; \mathcal{M}_t, \theta_{t-1}, D, D') \\
&\quad = \log \frac{p\left(\mathcal{M}_t(\theta_{t-1}, D) = \theta\right)}{p\left(\mathcal{M}_t(\theta_{t-1}, D') = \theta\right)} \\
&\quad = -\frac{1}{2\varsigma^2}\left(\left\|\theta - \mu_{\theta_{t-1},D}\right\|_2^2 - \left\|\theta - \mu_{\theta_{t-1},D'}\right\|_2^2\right) \\
&\quad = -\frac{1}{2\varsigma^2}\left(2\theta^\top(\mu_{\theta_{t-1},D'} - \mu_{\theta_{t-1},D}) + \left\|\mu_{\theta_{t-1},D}\right\|_2^2 - \left\|\mu_{\theta_{t-1},D'}\right\|_2^2\right) \\
&\quad = -\frac{1}{2\varsigma^2}\left(2(\theta - \mu_{\theta_{t-1},D})^\top(\mu_{\theta_{t-1},D'} - \mu_{\theta_{t-1},D}) - \left\|\mu_{\theta_{t-1},D}\right\|_2^2 + 2\mu_{\theta_{t-1},D}^\top \mu_{\theta_{t-1},D'} - \left\|\mu_{\theta_{t-1},D'}\right\|_2^2\right) \\
&\quad = -\frac{1}{2\varsigma^2}\left(2(\theta - \mu_{\theta_{t-1},D})^\top \Delta_{\theta_{t-1},D,D'} - \left\|\Delta_{\theta_{t-1},D,D'}\right\|_2^2\right),
\end{aligned} \tag{I.14}$$

where we introduce the shorthand $\Delta_{\theta_{t-1},D,D'} = \mu_{\theta_{t-1},D'} - \mu_{\theta_{t-1},D}$. We can now compute the

moment generating function of the privacy loss

$$
\begin{aligned}
&\alpha_{\mathcal{M}_t}(\lambda; \theta_{t-1}, D, D') \\
&\quad = \log \mathbb{E}_{\theta \sim \mathcal{M}_t(\theta_{t-1}, D)} \left[\exp(\lambda \gamma(\theta; \mathcal{M}_t, \theta_{t-1}, D, D')) \right] \\
&\quad = \log \left(\exp\left(\lambda \frac{\left\| \Delta_{\theta_{t-1}, D, D'} \right\|_2^2}{2\varsigma^2} \right) \mathbb{E}_{\theta \sim \mathcal{M}(\theta_{t-1}, D)} \left[\exp\left(-\lambda \frac{(\theta - \mu_{\theta_{t-1}, D})^\top \Delta_{\theta_{t-1}, D, D'}}{\varsigma^2} \right) \right] \right) \\
&\quad = \lambda \frac{\left\| \Delta_{\theta_{t-1}, D, D'} \right\|_2^2}{2\varsigma^2} + \log \mathbb{E}_{\theta \sim \mathcal{N}(\mu_{\theta_{t-1}, D}, \varsigma^2 I_p)} \left[\exp\left(-\lambda \frac{(\theta - \mu_{\theta_{t-1}, D})^\top \Delta_{\theta_{t-1}, D, D'}}{\varsigma^2} \right) \right] \\
&\quad = \lambda \frac{\left\| \Delta_{\theta_{t-1}, D, D'} \right\|_2^2}{2\varsigma^2} + \log \mathbb{E}_{\theta' \sim \mathcal{N}(0, I_p)} \left[\exp\left(-\lambda \frac{\theta'^\top \Delta_{\theta_{t-1}, D, D'}}{\varsigma} \right) \right] \\
&\quad = \lambda \frac{\left\| \Delta_{\theta_{t-1}, D, D'} \right\|_2^2}{2\varsigma^2} + \log \mathbb{E}_{\rho \sim \mathcal{N}(0,1)} \left[\exp\left(-\lambda \frac{\left\| \Delta_{\theta_{t-1}, D, D'} \right\|_2}{\varsigma} \rho \right) \right] \\
&\quad = \lambda \frac{\left\| \Delta_{\theta_{t-1}, D, D'} \right\|_2^2}{2\varsigma^2} + \lambda^2 \frac{\left\| \Delta_{\theta_{t-1}, D, D'} \right\|_2^2}{2\varsigma^2} \\
&\quad = \frac{\left\| \Delta_{\theta_{t-1}, D, D'} \right\|_2^2}{2\varsigma^2} \left(\lambda + \lambda^2 \right),
\end{aligned}
\tag{I.15}
$$

where we perform a change of variable in the fifth line, use the rotational invariance of the Gaussian distribution in the sixth line, and compute the moment generating function of a standard Gaussian on the seventh line. Thus, we can write

$$
\alpha_{\mathcal{M}_t}(\lambda) = \sup_{\theta_{t-1}, D \text{ adjacent with } D'} \alpha_{\mathcal{M}_t}(\lambda; \theta_{t-1}, D, D') = \frac{(\Delta_2 \mu)^2}{2\varsigma^2} \left(\lambda + \lambda^2 \right), \tag{I.16}
$$

where

$$
\Delta_2 \mu = \sup_{\theta_{t-1}, D \text{ adjacent with } D'} \left\| \Delta_{\theta_{t-1}, D, D'} \right\|_2 \tag{I.17}
$$

is the ℓ_2 sensitivity of the gradient updates in Algorithm 8.1. By Lemma I.2, we have that $\Delta_2 \mu \leq 2\eta C_{\mathsf{clip}}/n$. Thus, plugging $\varsigma = \sqrt{\eta} \frac{2C_{\mathsf{clip}}}{n} \sigma$ in (I.16), we readily get the desired result. □

Proof of Proposition 8.2. Algorithm 8.1 is a randomized algorithm $\mathcal{A}$ consisting of a sequence of independent adaptive mechanisms $\mathcal{M}_1, \ldots, \mathcal{M}_T$ such that, for all $t \in [T]$, by Lemma I.3, we have

$$
\alpha_{\mathcal{M}_t}(\lambda) \leq \frac{\eta}{2\sigma^2} \left(\lambda + \lambda^2 \right). \tag{I.18}
$$

Thus, by composability (see Theorem I.1), we have

$$
\alpha_{\mathcal{A}}(\lambda) \leq \frac{\eta T}{2\sigma^2} \left(\lambda + \lambda^2 \right). \tag{I.19}
$$

Set $\delta \in (0, 1)$ and $\varepsilon \in (0, 8 \log(1/\delta))$, and suppose that

$$
\frac{\eta T}{2\sigma^2} \leq \frac{\varepsilon^2}{16 \log(1/\delta)}. \tag{I.20}
$$

Then, considering $\lambda > 0$, we have that

$$\begin{aligned}\exp\left(\alpha_{\mathcal{A}}(\lambda) - \lambda\varepsilon\right) &\leq \exp\left(\frac{\varepsilon^2}{16\log(1/\delta)}\left(\lambda + \lambda^2\right) - \lambda\varepsilon\right) \\ &\leq \exp\left(\frac{\varepsilon^2}{16\log(1/\delta)}\lambda^2 - \left(1 - \frac{\varepsilon}{16\log(1/\delta)}\right)\lambda\varepsilon\right) \\ &\leq \exp\left(\frac{\varepsilon^2}{16\log(1/\delta)}\lambda^2 - \frac{\lambda\varepsilon}{2}\right).\end{aligned} \tag{I.21}$$

Setting $\lambda^* = 4\log(1/\delta)/\varepsilon$ we get

$$\begin{aligned}\exp\left(\alpha_{\mathcal{A}}(\lambda^*) - \lambda^*\varepsilon\right) &\leq \exp\left(\frac{\varepsilon^2}{16\log(1/\delta)}\lambda^{*2} - \frac{\lambda^*\varepsilon}{2}\right) \\ &= \exp\left(\log(1/\delta) - 2\log(1/\delta)\right) \\ &= \exp\left(-\log(1/\delta)\right) = \delta.\end{aligned} \tag{I.22}$$

Then, by the tail bound in Theorem I.1, we have that $\mathcal{A}$ is (ε, δ)-differentially private. In our argument, we consider $\delta \in (0, 1)$, $\varepsilon \in (0, 8\log(1/\delta))$, and the inequality in (I.20), which can be rewritten as

$$\sigma \geq \sqrt{\eta T}\frac{\sqrt{8\log(1/\delta)}}{\varepsilon}, \tag{I.23}$$

which readily gives the desired result. □

We now move to the proof of Proposition 8.3, and then present the auxiliary Proposition I.4 that will be useful in the discussion on the SDE defined in (8.8). Next, in Proposition I.5 we formalize the convergence of Algorithm 8.1 in the limit of small learning rates $\eta \to 0$. To conclude, we state the auxiliary Lemma I.6, useful for the final proof of Proposition 8.4.

Proof of Proposition 8.3. At every iteration t of Algorithm 8.1, we have

$$\begin{aligned}g_{C_{\mathsf{clip}}}(x_i, y_i, \theta_{t-1}) &= g(x_i, y_i, \theta_{t-1}) / \max\left(1, \frac{\|g(x_i, y_i, \theta_{t-1})\|_2}{C_{\mathsf{clip}}}\right) \\ &= \frac{\varphi(x_i)\ell'(\varphi(x_i)^\top\theta_{t-1} - y_i)}{\max\left(1, \frac{\|\varphi(x_i)\|_2\left|\ell'\left(\varphi(x_i)^\top\theta_{t-1} - y_i\right)\right|}{C_{\mathsf{clip}}}\right)} \\ &= \varphi(x_i)\ell'_{i,C_{\mathsf{clip}}}\left(\varphi(x_i)^\top\theta_{t-1} - y_i\right) \\ &= \nabla_\theta \ell_{i,C_{\mathsf{clip}}}\left(\varphi(x_i)^\top\theta_{t-1} - y_i\right). \quad \square\end{aligned} \tag{I.24}$$

Proposition I.4. *Let $\ell : \mathbb{R} \to \mathbb{R}$ be a differentiable function with Lipschitz-continuous derivative. Then, for all $\theta \in \mathbb{R}^p$, we have*

$$\left\|\nabla\mathcal{L}_{C_{\mathsf{clip}}}(\theta)\right\|_2 \leq C_{\mathsf{clip}}. \tag{I.25}$$

Furthermore, there exists a constant K such that, for all $\theta, \theta' \in \mathbb{R}^p$, we have

$$\left\|\nabla\mathcal{L}_{C_{\mathsf{clip}}}(\theta) - \nabla\mathcal{L}_{C_{\mathsf{clip}}}(\theta')\right\|_2 \leq K\left\|\theta - \theta'\right\|_2, \tag{I.26}$$

with $K \leq L\sum_{i=1}^n \|\varphi(x_i)\|_2^2/n$, where L is the Lipschitz constant of ℓ'.

Proof. Following the definition in (8.6), we have that, for all $i \in [n]$,

$$\nabla_\theta \ell_{i,C_{\text{clip}}}\left(\varphi(x_i)^\top\theta - y_i\right) = \nabla_\theta \ell\left(\varphi(x_i)^\top\theta - y_i\right) / \max\left(1, \frac{\left\|\nabla_\theta \ell\left(\varphi(x_i)^\top\theta - y_i\right)\right\|_2}{C_{\text{clip}}}\right), \tag{I.27}$$

which readily gives

$$\left\|\nabla_\theta \ell_{i,C_{\text{clip}}}\left(\varphi(x_i)^\top\theta - y_i\right)\right\|_2 \leq C_{\text{clip}}. \tag{I.28}$$

Then, we can conclude

$$\begin{aligned}\left\|\nabla \mathcal{L}_{C_{\text{clip}}}(\theta)\right\|_2 &= \left\|\nabla_\theta \frac{1}{n}\sum_{i=1}^n \ell_{i,C_{\text{clip}}}(\varphi(x_i)^\top\theta - y_i)\right\|_2 \\ &\leq \frac{1}{n}\sum_{i=1}^n \left\|\nabla_\theta \ell_{i,C_{\text{clip}}}(\varphi(x_i)^\top\theta - y_i)\right\|_2 \\ &\leq C_{\text{clip}}.\end{aligned} \tag{I.29}$$

For the second part of the claim, following a similar argument as before and exploiting the chain rule, it is sufficient to show that, for every $i \in [n]$, $\ell_{i,C_{\text{clip}}}$ has Lipschitz-continuous derivative. For all $z \in \mathbb{R}$ such that $\ell'(z) > 0$, we can write $\ell'_{i,C_{\text{clip}}}(z) = \min\left(\ell'(z), C_{\text{clip}}/\left\|\varphi(x_i)\right\|_2\right)$. Then, if ℓ' is Lipschitz-continuous, also $\ell'_{i,C_{\text{clip}}}$ is, as it can be written as composition of Lipschitz-continuous functions. A similar argument holds for all $z \in \mathbb{R}$ such that $\ell'(z) < 0$. Furthermore, for all z such that $|\ell'(z)| < C_{\text{clip}}/\left\|\varphi(x_i)\right\|_2$, we simply have $\ell'_{i,C_{\text{clip}}} = \ell'$. Thus, to prove that $\ell'_{i,C_{\text{clip}}}$ is Lipschitz-continuous, it suffices that ℓ' is Lipschitz-continuous, as assumed by the proposition. Notice that this argument also proves that the Lipschitz constant of $\ell'_{i,C_{\text{clip}}}$ is smaller or equal than the Lipschitz constant L of ℓ'. Thus, we have

$$\begin{aligned}\left\|\nabla \mathcal{L}_{C_{\text{clip}}}(\theta) - \nabla \mathcal{L}_{C_{\text{clip}}}(\theta')\right\|_2 &= \left\|\nabla \frac{1}{n}\sum_{i=1}^n \ell_{i,C_{\text{clip}}}(\varphi(x_i)^\top\theta - y_i) - \nabla \frac{1}{n}\sum_{i=1}^n \ell_{i,C_{\text{clip}}}(\varphi(x_i)^\top\theta' - y_i)\right\|_2 \\ &= \left\|\frac{1}{n}\sum_{i=1}^n \varphi(x_i)\left(\ell'_{i,C_{\text{clip}}}(\varphi(x_i)^\top\theta - y_i) - \ell'_{i,C_{\text{clip}}}(\varphi(x_i)^\top\theta' - y_i)\right)\right\|_2 \\ &\leq \frac{1}{n}\sum_{i=1}^n \left\|\varphi(x_i)\right\|_2 \left|\ell'_{i,C_{\text{clip}}}(\varphi(x_i)^\top\theta - y_i) - \ell'_{i,C_{\text{clip}}}(\varphi(x_i)^\top\theta' - y_i)\right| \\ &\leq \frac{L}{n}\sum_{i=1}^n \left\|\varphi(x_i)\right\|_2^2 \left\|\theta - \theta'\right\|_2,\end{aligned}$$

which proves the last statement. □

Due to Proposition I.4, the SDE in (8.8) satisfies the assumptions A1-4 in Section 4.5 of Kloeden and Platen [2011]. This means that (8.8) admits a unique strong solution $\Theta(t)$ (see Theorem 4.5.3 in Kloeden and Platen [2011]). Furthermore, we also have that the assumptions in Equations (1.5), (1.6), (1.7) of Menozzi et al. [2021] are verified, implying that at every fixed time $t > 0$, the probability density function $p(\Theta(t))$ is continuous with respect to the coordinate θ (see Theorem 1.2 in Menozzi et al. [2021]). This property will be used later in Lemma I.6 and Theorem 8.4.

In Section 8.3, we have mentioned that (8.7) is the Euler-Maruyama discretization of (8.8). This can be formalized defining a family of random variables $\{\theta_T^1, \theta_T^2, \dots, \}$, where θ_T^j is the

output of Algorithm 8.1 with learning rate $\eta_j := \eta/j$ and number of iterations $T_j := jT$, with j being a positive integer. Algorithm 8.1 is defined such that, when performing the t-th iteration, we introduce the Gaussian random variable

$$\frac{2C_{\text{clip}}}{n}\sigma\left(B(t\eta/j) - B((t-1)\eta/j)\right), \tag{I.30}$$

where $B(t)$ is the standard p-dimensional Wiener process in (8.8). This is coherent with the definition of Algorithm 8.1, since these random variables are distributed as

$$\frac{2C_{\text{clip}}}{n}\sigma\mathcal{N}(0, \eta I/j) = \sqrt{\eta_j}\frac{2C_{\text{clip}}}{n}\sigma\mathcal{N}(0, I), \tag{I.31}$$

and they are independent with each other. Since (8.8) respects all the necessary assumptions in Theorem 10.2.2 in Kloeden and Platen [2011], the following result holds:

Proposition I.5. *Let $\tau = \eta T$. Then, for every $h > 0$, there exists j^*, such that for all $j > j^*$, we have*

$$\mathbb{E}\left[\left\|\theta_T^j - \Theta(\tau)\right\|_2\right] < h. \tag{I.32}$$

Proposition I.5 formalizes in which sense the SDE in (8.8) is a continuous limit of Algorithm 8.1. This is done through the family of algorithms with progressively smaller learning rates and larger number of iterations. It is important to notice that all these algorithms provide the same DP guarantee, as $\eta_j T_j = \eta T$ for every j, and the result of Proposition 8.2 depends on learning rate and number of iterations only through their product ηT.

Lemma I.6. *We have that, for any open ball $S \subseteq \mathbb{R}^p$,*

$$\lim_{j\to\infty} \mathbb{P}\left(\theta_T^j \in S\right) = \mathbb{P}\left(\Theta_\tau \in S\right). \tag{I.33}$$

Proof. By contradiction, let the thesis be false. Then, there exists a ball $S(r_0, r)$ with center r_0 and radius r s.t. there is $h^* > 0$ and j arbitrarily large that satisfy

$$\left|\mathbb{P}\left(\theta_T^j \in S(r_0, r)\right) - \mathbb{P}\left(\Theta_\tau \in S(r_0, r)\right)\right| > h^*. \tag{I.34}$$

We first suppose that we have

$$\mathbb{P}\left(\theta_T^j \in S(r_0, r)\right) > \mathbb{P}\left(\Theta_\tau \in S(r_0, r)\right) + h^*, \tag{I.35}$$

for j arbitrarily large. Then, there is a sequence of events ω_j s.t. $\mathbb{P}(\omega_j) > h^*$ and

$$\theta_T^j(\omega_j) \in S(r_0, r), \qquad \Theta_\tau(\omega_j) \notin S(r_0, r). \tag{I.36}$$

As mentioned previously in this section, the law $p(\Theta_\tau)$ is continuous, and therefore bounded in the closed set $S(r_0, r')$ for any r'. This implies that there exists $r^* > 0$, such that

$$\mathbb{P}\left(\Theta_\tau \notin S(r_0, r) \text{ and } \Theta_\tau \in S(r_0, r + r^*)\right) < \frac{h^*}{2}, \tag{I.37}$$

which in turn implies that there is a sequence of events ω_j^*, such that $\mathbb{P}\left(\omega_j^*\right) > h^*/2$ and

$$\theta_T^j(\omega_j^*) \in S(r_0, r), \qquad \Theta_\tau(\omega_j^*) \notin S(r_0, r + r^*). \tag{I.38}$$

Note that, for all such events, by triangle inequality, we have

$$\begin{aligned}\left\|\theta_T^j(\omega_j^*) - \Theta_\tau(\omega_j^*)\right\|_2 &\geq \left\|\Theta_\tau(\omega_j^*) - r_0\right\|_2 - \left\|\theta_T^j(\omega_j^*) - r_0\right\|_2 \\ &\geq (r + r^* - r_0) - (r - r_0) \\ &= r^*,\end{aligned} \tag{I.39}$$

which implies

$$\mathbb{E}\left[\left\|\theta_T^j - \Theta_\tau\right\|_2\right] \geq \mathbb{P}\left(\omega_j^*\right)\mathbb{E}\left[\left\|\theta_T^j - \Theta_\tau\right\|_2 \middle| \omega_j^*\right] \geq \frac{h^* r^*}{2}. \tag{I.40}$$

This last equation holds for arbitrarily large j, which provides the desired contradiction with Proposition I.5.

If instead of (I.35) we have

$$\mathbb{P}\left(\Theta_\tau \in S(r_0, r)\right) > \mathbb{P}\left(\theta_T^j \in S(r_0, r)\right) + h^*, \tag{I.41}$$

for j arbitrarily large, then the argument follows in the same way. In fact, we exploit the continuity of the law $p(\Theta_\tau)$ to state that there exists $r^* > 0$, such that

$$\mathbb{P}\left(\Theta_\tau \notin S(r_0, r - r^*) \text{ and } \Theta_\tau \in S(r_0, r)\right) < \frac{h^*}{2}, \tag{I.42}$$

thus giving a sequence of events ω_j^*, such that $\mathbb{P}\left(\omega_j\right) > h^*/2$ and

$$\theta_T^j(\omega_j^*) \notin S(r_0, r), \qquad \Theta_\tau(\omega_j^*) \in S(r_0, r - r^*). \tag{I.43}$$

At this point, we can again apply the triangle inequality to obtain the same contradiction as in (I.40). □

Proof of Proposition 8.4. The assumption on Σ guarantees, due to Theorem 8.2, that for all $j > 0$, the discretizations θ_T^j of the SDE (8.8) are, for any $\delta \in (0, 1)$ and $\varepsilon \in (0, 8\log(1/\delta))$, (ε, δ)-differentially private, since

$$\sigma = \frac{n}{2C_{\text{clip}}}\Sigma \geq \sqrt{\tau}\frac{\sqrt{8\log(1/\delta)}}{\varepsilon} = \sqrt{\eta_j T_j}\frac{\sqrt{8\log(1/\delta)}}{\varepsilon}, \tag{I.44}$$

where the first step is due to (8.9).

By contradiction, suppose that Θ_τ is not (ε, δ)-DP. This means that there exists a point in the parameters space $\bar{\theta}$ such that, on adjacent datasets D and D', we have

$$p\left(\Theta_\tau(D) = \bar{\theta}\right) > e^\varepsilon p\left(\Theta_\tau(D') = \bar{\theta}\right) + \delta. \tag{I.45}$$

Since both the laws in the previous equation are continuous, we also have that there exists an open ball S centered in $\bar{\theta}$, such that

$$\mathbb{P}\left(\Theta_\tau(D) \in S\right) > e^\varepsilon \mathbb{P}\left(\Theta_\tau(D') \in S\right) + \delta, \tag{I.46}$$

which we rewrite as

$$\mathbb{P}\left(\Theta_\tau(D) \in S\right) = e^\varepsilon \mathbb{P}\left(\Theta_\tau(D') \in S\right) + \delta + h, \tag{I.47}$$

for some $h > 0$. Now, by Lemma I.6, we have that there exists j^* large enough s.t.

$$\left|\mathbb{P}\left(\theta_T^{j^*}(D) \in S\right) - \mathbb{P}\left(\Theta_\tau(D) \in S\right)\right| < \frac{h}{2}, \tag{I.48}$$

and

$$\left|\mathbb{P}\left(\theta_T^{j^*}(D') \in S\right) - \mathbb{P}\left(\Theta_\tau(D') \in S\right)\right| < \frac{he^{-\varepsilon}}{2}. \tag{I.49}$$

Hence, putting together (I.47), (I.48), and (I.49), we have

$$\begin{aligned}
\mathbb{P}\left(\theta_T^{j^*}(D) \in S\right) &> \mathbb{P}\left(\Theta_\tau(D) \in S\right) - \frac{h}{2} \\
&= e^{\varepsilon}\mathbb{P}\left(\Theta_\tau(D') \in S\right) + \delta + h - \frac{h}{2} \\
&> e^{\varepsilon}\left(\mathbb{P}\left(\theta_T^{j^*}(D') \in S\right) - \frac{he^{-\varepsilon}}{2}\right) + \delta + h - \frac{h}{2} \\
&= e^{\varepsilon}\mathbb{P}\left(\theta_T^{j^*}(D') \in S\right) + \delta.
\end{aligned} \tag{I.50}$$

The previous equation implies that the discretization $\theta_T^{j^*}$ is not (ε, δ)-DP, therefore contradicting the thesis of Theorem 8.2. This provides the desired result. □

I.3 Auxiliary Lemmas

This section uses the following additional notation. We denote by $\mathbf{1} : \mathbb{R} \to \{0, 1\}$ the indicator function, *i.e.*, $\mathbf{1}(z) = 1$ if $z > 0$, and 0 otherwise. Given a matrix A, we indicate with $\sigma_{\min}(A) = \sqrt{\lambda_{\min}\left(A^\top A\right)}$ its smallest singular value. Given a p.s.d. matrix A, $\lambda_j(A)$ denotes its j-th eigenvalue sorted in non-increasing order ($\lambda_{\max}(A) = \lambda_1(A) \geq \lambda_2(A) \geq \ldots \geq \lambda_s(A) = \lambda_{\min}(A)$). Given two matrices $A \in \mathbb{R}^{s \times s_1}$ and $B \in \mathbb{R}^{s \times s_2}$, we denote by $A * B = [(A_{1:} \otimes B_{1:}), \ldots, (A_{s:} \otimes B_{s:})]^\top \in \mathbb{R}^{s \times s_1 s_2}$ their row-wise Kronecker product (also known as Khatri-Rao product), and $A_{j:}$ denotes the j-th row of A. Given a natural number l, we denote $A^{*l} = A * A^{*(l-1)}$, with $A^{*1} = A$. We remark that $\left[\left(A^{*l}\right)\left(A^{*l}\right)^\top\right]_{ij} = \left(\left[AA^\top\right]_{ij}\right)^l$. Given $A \in \mathbb{R}^{s \times s}$ and $B \in \mathbb{R}^{s \times s}$, we use the notation $A \preceq B$ ($B \succeq A$) to indicate that $B - A$ is p.s.d. As in Section 8.6, we indicate with $P_\Lambda \in \mathbb{R}^{p \times p}$ the projector on the space spanned by the eigenvectors associated with the d largest eigenvalues of $\Phi^\top\Phi$, and we condition on the high probability event described by Lemma I.11, which guarantees a spectral gap in $\Phi^\top\Phi$ (and, hence, the well-posedness of P_Λ).

Lemma I.7. *Let Assumption 8.5 hold, and let $x \sim \mathcal{P}_X$. Then, we have that*

$$\|\mathbb{E}[x]\|_2 = O(1), \tag{I.51}$$

and

$$\left\|\mathbb{E}[xx^\top]\right\|_{\mathsf{op}} = O(1). \tag{I.52}$$

Proof. The first statement is a direct consequence of x being sub-Gaussian. In fact, this implies

$$\mathbb{E}\left[\left(x^\top \frac{\mathbb{E}[x]}{\|\mathbb{E}[x]\|_2}\right)^2\right] = O(1), \tag{I.53}$$

as the second moment of sub-Gaussian random variables is bounded. Thus, we have

$$\|\mathbb{E}[x]\|_2^2 = \mathbb{E}\left[x^\top \frac{\mathbb{E}[x]}{\|\mathbb{E}[x]\|_2}\right]^2 \leq \mathbb{E}\left[\left(x^\top \frac{\mathbb{E}[x]}{\|\mathbb{E}[x]\|_2}\right)^2\right] = O(1), \tag{I.54}$$

which provides the desired result.

For the second statement, we have, for every $u \in \mathbb{R}^d$ such that $\|u\|_2 = 1$,

$$u^\top \mathbb{E}\left[xx^\top\right] u = \mathbb{E}\left[\left(x^\top u\right)^2\right] = O(1), \tag{I.55}$$

where the second step holds as the second moment of sub-Gaussian random variables is bounded. Since $\mathbb{E}\left[xx^\top\right]$ is p.s.d., its operator norm is $\sup_{\|u\|_2=1} u^\top \mathbb{E}\left[xx^\top\right] u$, and the second statement readily follows. □

Lemma I.8. *Let Assumption 8.5 hold, and let $d = o(n)$ and $d = o(p)$. Then, we have that*

$$\lambda_{\min}\left(X^\top X\right) = \Theta(n), \tag{I.56}$$

and

$$\|X\|_{\mathsf{op}} = O\left(\sqrt{n}\right), \tag{I.57}$$

with probability at least $1 - 2\exp(-cd)$ over X, where c is an absolute constant.

Also, we have that

$$\lambda_{\min}\left(V^\top V\right) = \Theta\left(\frac{p}{d}\right), \tag{I.58}$$

and

$$\|V\|_{\mathsf{op}} = O\left(\sqrt{\frac{p}{d}}\right), \tag{I.59}$$

with probability at least $1 - 2\exp(-cd)$ over V.

Proof. Notice that $X \in \mathbb{R}^{n\times d}$ is a matrix with i.i.d. sub-Gaussian rows with second moment matrix $\Psi = \mathbb{E}_{x\sim\mathcal{P}_X}\left[xx^\top\right] \in \mathbb{R}^{d\times d}$. Thus, by Remark 5.40 in Vershynin [2012], we have that

$$\left\|X^\top X - n\Psi\right\|_{\mathsf{op}} = O\left(n\frac{d}{n}\right) = o(n), \tag{I.60}$$

with probability at least $1 - 2\exp(-c_2 d)$. Then, conditioning on this high probability event, by Weyl's inequality, we have

$$\lambda_{\min}\left(X^\top X\right) \leq \left\|X^\top X\right\|_{\mathsf{op}} \leq \|n\Psi\|_{\mathsf{op}} + \left\|X^\top X - n\Psi\right\|_{\mathsf{op}} = O(n), \tag{I.61}$$

where the last step is a consequence of Lemma I.7. This proves the upper bound in the first statement as well as the second statement. Furthermore, again by Weyl's inequality, we have

$$\lambda_{\min}\left(X^\top X\right) \geq \lambda_{\min}(n\Psi) - \left\|X^\top X - n\Psi\right\|_{\mathsf{op}} = \Omega(n), \tag{I.62}$$

where the last step is a consequence of $\lambda_{\min}(\Psi) = \Omega(1)$, true by Assumption 8.5. Merging this with (I.61) concludes the proof of the first statement.

The third statement can be proven in the same exact way, considering $\sqrt{d}V \in \mathbb{R}^{p\times d}$, a matrix with i.i.d. standard Gaussian (and therefore sub-Gaussian) rows with second moment matrix equal to the identity $I \in \mathbb{R}^{d\times d}$. Finally, the fourth statement is a consequence of Theorem 4.4.5 of Vershynin [2018] and the scaling $d = o(p)$, and it holds with probability $1 - 2\exp(-c_3 d)$. This concludes the proof and provides the desired results. □

Lemma I.9. *Let Assumptions 8.5 and 8.6 hold, and let* $n \log^3 n = o\left(d^{3/2}\right)$ *and* $n = \omega(d)$. *Then with probability at least* $1 - 2\exp\left(-c\log^2 n\right)$ *over* X, *all the following hold*

$$\lambda_d(\mathbb{E}_V[K]) = \Omega\left(\frac{pn}{d}\right), \tag{I.63}$$

$$\lambda_{d+1}(\mathbb{E}_V[K]) = O\left(p\right), \tag{I.64}$$

$$\lambda_n(\mathbb{E}_V[K]) = \lambda_{\min}\left(\mathbb{E}_V[K]\right) = \Omega(p), \tag{I.65}$$

where c *is an absolute constant.*

Proof. As $K_{ij} = \varphi(x_i)^\top \varphi(x_j)$, we have

$$\mathbb{E}_V[K]_{ij} = \sum_{k=1}^{p} \mathbb{E}_{v_k}[\phi(v_k^\top x_i)\phi(v_k^\top x_j)] = p\,\mathbb{E}_v[\phi(v^\top x_i)\phi(v^\top x_j)], \tag{I.66}$$

where in the last step we exploited that the v_k-s are identically distributed, and introduced the shorthand v to indicate a random variable distributed as v_1. Since $\|x_i\| = \sqrt{d}$ for all $i \in [n]$, and $v \sim \mathcal{N}(0, 1/d)$, we have that $\rho_1 := v^\top x_i$ and $\rho_2 := v^\top x_j$ are two standard Gaussian variables with correlation $x_i^\top x_j/d$. Thus, exploiting the Hermite expansion of ϕ, we can write

$$\mathbb{E}_V[K]_{ij} = p\sum_{l=0}^{+\infty} \mu_l^2 \frac{\left(x_i^\top x_j\right)^l}{d^l} = p\sum_{l=0}^{\infty} \mu_l^2 \frac{\left[\left(X^{*l}\right)\left(X^{*l}\right)^\top\right]_{ij}}{d^l}, \tag{I.67}$$

where μ_l is the l-th Hermite coefficient of ϕ. Since ϕ is such that $\mu_l = 0$ for $l = \{0, 2\}$, $\mu_1 \neq 0$, and there exists $l \geq 3$ such that $\mu_l \neq 0$ (since ϕ is non-linear), we can write the previous sum with the two non-0 terms

$$\mathbb{E}_V[K]_{ij} = \frac{\mu_1^2 p}{d} XX^\top + p\sum_{l=3}^{+\infty} \mu_l^2 \frac{\left[\left(X^{*l}\right)\left(X^{*l}\right)^\top\right]_{ij}}{d^l}. \tag{I.68}$$

We analyze the terms of the sum separately. By Lemma I.8 we have that, with probability at least $1 - 2\exp\left(-c_1 d\right)$ over X,

$$\lambda_{\min}\left(X^\top X\right) = \Theta(n). \tag{I.69}$$

As $X \in \mathbb{R}^{n\times d}$, with $d < n$, this means that (conditioning on this high probability event) $XX^\top$ is a matrix with rank equal to d, and that its first d eigenvalues (when sorted in non-increasing order) are all $\Omega(n)$. This implies that

$$M_1 := \frac{\mu_1^2 p}{d} XX^\top \tag{I.70}$$

has the first d eigenvalues of order $\Omega(pn/d)$, and the remaining $n - d$ equal to 0.

We now consider the second term of (I.68):

$$M_2 = p\sum_{l=3}^{\infty} \mu_l^2 \frac{\left[\left(X^{*l}\right)\left(X^{*l}\right)^\top\right]_{ij}}{d^l}. \tag{I.71}$$

Let us define

$$\tilde{M}_2 = p\left(\sum_{l=3}^{\infty} \mu_l^2\right) I = CpI, \tag{I.72}$$

corresponding to the diagonal elements of M_2, where C is a constant depending only on ϕ. Notice that C is strictly positive, as ϕ is non-linear. For $i \neq j$, since x_i is sub-Gaussian and independent from x_j, we have

$$\mathbb{P}\left(\left|x_i^\top x_j\right| > \log n\sqrt{d}\right) < 2\exp\left(-c_2 \log^2 n\right). \tag{I.73}$$

Performing a union bound we also get

$$\mathbb{P}\left(\max_{i,j}\left|x_i^\top x_j\right| > \log n\sqrt{d}\right) < 2n^2 \exp\left(-c_2 \log^2 n\right) < 2\exp\left(-c_3 \log^2 n\right). \tag{I.74}$$

This implies that

$$\begin{aligned}
\left\|M_2 - \tilde{M}_2\right\|_{\mathsf{op}} &\leq \left\|M_2 - \tilde{M}_2\right\|_F \\
&= p\left\|\sum_{l=3}^{\infty} \mu_l^2 \left(\frac{\left(X^{*l}\right)\left(X^{*l}\right)^\top}{d^l} - I\right)\right\|_F \\
&\leq p\sum_{l=3}^{\infty} \mu_l^2 \left\|\frac{\left(X^{*l}\right)\left(X^{*l}\right)^\top}{d^l} - I\right\|_F \\
&\leq p\sum_{l=3}^{\infty} \mu_l^2 \sqrt{n^2\left(\frac{\max_{i,j}\left|x_i^\top x_j\right|^l}{d^l}\right)^2} \\
&\leq p\sqrt{n^2\left(\frac{\max_{i,j}\left|x_i^\top x_j\right|^3}{d^3}\right)^2}\sum_{l=3}^{\infty} \mu_l^2 \\
&= O\left(p\frac{n\log^3 n}{d^{3/2}}\right) \\
&= o(p).
\end{aligned} \tag{I.75}$$

Merging (I.72) and (I.75) together, a standard application of Weyl's inequality gives

$$\lambda_{\max}\left(M_2\right) = \Theta(p), \qquad \lambda_{\min}\left(M_2\right) = \Theta(p), \tag{I.76}$$

with probability at least $1 - 2\exp\left(-c_3 \log^2 n\right)$. Thus, recalling that $\lambda_i(\cdot)$ denotes the i-th eigenvalue of a matrix sorted in non-increasing order and applying Weyl's inequality, we have that

$$\begin{aligned}
&\lambda_i\left(M_1 + M_2\right) \geq \lambda_i\left(M_1\right) - \lambda_{\max}\left(M_2\right) = \Omega(pn/d), \qquad \text{for } i \in [d], \\
&\lambda_i\left(M_1 + M_2\right) \leq \lambda_i\left(M_1\right) + \lambda_{\max}\left(M_2\right) = O\left(p\right), \qquad \text{for } d < i \leq n,
\end{aligned} \tag{I.77}$$

where in both lines we used our argument in (I.70) and (I.76). Since $\mathbb{E}_V[K] = M_1 + M_2$, we have proved that the spectrum of $\mathbb{E}_V[K]$ has a gap, as

$$\lambda_d\left(\mathbb{E}_V[K]\right) = \Omega\left(\frac{pn}{d}\right) = \omega(p), \tag{I.78}$$

and

$$\lambda_{d+1}\left(\mathbb{E}_V[K]\right) = O\left(p\right). \tag{I.79}$$

This proves (I.63) and (I.64). To show (I.65), it suffices to note that

$$\lambda_{\min}\left(\mathbb{E}_V[K]\right) \geq \lambda_{\min}\left(M_2\right) = \Theta(p), \tag{I.80}$$

where the first step is true since M_1 is p.s.d., and the second step follows from (I.76). □

Lemma I.10. *Let Assumptions 8.5 and 8.6 hold, and let* $n = O\left(p/\log^4 p\right)$, $n\log^3 n = o\left(d^{3/2}\right)$ *and* $n = \omega(d)$. *Then, we have that*

$$\left\|\mathbb{E}_V[K]^{-1/2}\left(K - \mathbb{E}_V[K]\right)\mathbb{E}_V[K]^{-1/2}\right\|_{\mathsf{op}} = O\left(\sqrt{\frac{n}{p}}\log n \log p\right), \tag{I.81}$$

with probability at least $1 - 2\exp\left(-c\log^2 n\right)$ *over* X *and* V, *where* c *is an absolute constant.*

Proof. Consider the truncated function

$$\bar{\phi}(z) := \phi(z)\mathbf{1}\left(|z| \leq \frac{\log p}{L}\right), \tag{I.82}$$

where L is the Lipschitz constant of ϕ. Define also the truncated kernel $\bar{K}$ accordingly

$$\bar{K}_{ij} := \bar{\phi}(Vx_i)^\top \bar{\phi}(Vx_j). \tag{I.83}$$

We now compare $\mathbb{E}_V\left[\bar{K}\right]$ and $\mathbb{E}_V[K]$. Let $v \sim \mathcal{N}(0, I/d)$ and define the random variable

$$E_{ij} := \mathbf{1}\left(|v^\top x_i| > \frac{\log p}{L} \text{ or } |v^\top x_j| > \frac{\log p}{L}\right). \tag{I.84}$$

As E_{ij} is an indicator, $E_{ij} = E_{ij}^2$ and

$$\mathbb{P}_V(E_{ij} = 1) \leq \mathbb{P}_V\left(|v^\top x_i| > \frac{\log p}{L}\right) + \mathbb{P}_V\left(|v^\top x_j| > \frac{\log p}{L}\right) \leq 2\exp\left(-c_1\log^2 p\right), \tag{I.85}$$

where the second step holds as $v^\top x_j \sim \mathcal{N}(0,1)$ and $v^\top x_i \sim \mathcal{N}(0,1)$ in the probability space of V. Thus, a union bound over i and j gives

$$K = \bar{K}, \tag{I.86}$$

with probability at least $1 - 2n^2\exp\left(-c_1\log^2 p\right) \geq 1 - 2\exp\left(-c_2\log^2 n\right)$ over V. Furthermore, we can write

$$\begin{aligned}
\left|\mathbb{E}_V\left[\bar{K}_{ij}\right] - \mathbb{E}_V[K_{ij}]\right| &= p\left|\mathbb{E}_v\left[\phi(v^\top x_i)\phi(v^\top x_j)\, E_{ij}\right]\right| \\
&\leq p\,\mathbb{E}_v\left[\left(\phi(v^\top x_i)\phi(v^\top x_j)\right)^2\right]^{1/2}\mathbb{E}_v\left[E_{ij}^2\right]^{1/2} \\
&\leq p\,\mathbb{E}_v\left[\left(\phi(v^\top x_i)\right)^4\right]^{1/2}\mathbb{P}_V(E_{ij}=1)^{1/2} \\
&\leq 2p\,C_1^{1/2}\exp\left(-\frac{c_1\log^2 p}{2}\right) \\
&\leq 2p\exp\left(-c_3\log^2 p\right).
\end{aligned} \tag{I.87}$$

Here, the second and third line follow from Cauchy-Schwartz inequality; the fourth line is a consequence of (I.85) and $\phi(v^\top x_i)$ being a sub-Gaussian random variable ($\phi(z)$ is Lipschitz) and thus with bounded fourth moment (see Equation 2.11 in Vershynin [2018]). This result holds for any $i, j \in [n]$, and therefore implies

$$\begin{aligned}\left\|\mathbb{E}_V\left[\bar{K}\right] - \mathbb{E}_V[K]\right\|_{\mathsf{op}} &\leq \left\|\mathbb{E}_V\left[\bar{K}\right] - \mathbb{E}_V[K]\right\|_F \\ &\leq 2pn\exp\left(-c_3\log^2 p\right) \\ &\leq 2\exp\left(-c_4\log^2 p\right).\end{aligned} \tag{I.88}$$

Then, by Weyl's inequality, we have

$$\lambda_{\min}\left(\mathbb{E}_V\left[\bar{K}\right]\right) \geq \lambda_{\min}\left(\mathbb{E}_V\left[K\right]\right) - \left\|\mathbb{E}_V\left[\bar{K}\right] - \mathbb{E}_V[K]\right\|_{\mathsf{op}} = \Omega(p), \tag{I.89}$$

where the last step follows from Lemma I.9 and it holds with probability at least $1 - 2\exp\left(-c_5\log^2 n\right)$ over X.

We now prove the bound in (I.81) on the truncated kernel $\bar{K}$. Let us then define the matrix $H_k \in \mathbb{R}^{n\times n}$ as

$$H_k = \mathbb{E}_V[\bar{K}]^{-1/2}\bar{\phi}(Xv_k)\bar{\phi}(Xv_k)^\top\mathbb{E}_V[\bar{K}]^{-1/2}. \tag{I.90}$$

This definition requires $\mathbb{E}_V[\bar{K}]$ to be invertible, and it is therefore conditioned on the event described by (I.89). We will condition on this high probability event over X until the end of the proof. Then, we have

$$\mathbb{E}_V[\bar{K}]^{-1/2}\left(\bar{K} - \mathbb{E}_V[\bar{K}]\right)\mathbb{E}_V[\bar{K}]^{-1/2} = \sum_{k=1}^{p} H_k - \mathbb{E}_V\left[H_k\right]. \tag{I.91}$$

Note that there exists a constant C such that

$$\begin{aligned}\sup_{v_k}\left\|H_k - \mathbb{E}_V[H_k]\right\|_{\mathsf{op}} \leq 2\sup_{v_k}\left\|H_k\right\|_{\mathsf{op}} &\leq 2\left\|\mathbb{E}_V[\bar{K}]^{-1/2}\right\|_{\mathsf{op}}^2\sup_{v_k}\left\|\bar{\phi}(Xv_k)\right\|_2^2 \\ &\leq C\frac{\sup_{v_k}\left\|\bar{\phi}(Xv_k)\right\|_2^2}{p} \leq C\frac{n}{p}\log^2 p,\end{aligned} \tag{I.92}$$

where the first step is true because of Jensen and triangle inequality, the third step is true because of (I.89), and the last step holds because, for every $i \in [n]$,

$$\left|\bar{\phi}(v_k^\top x_i)\right| \leq |\phi(0)| + L\frac{\log p}{L} = |\phi(0)| + \log p \leq C_3\log p, \tag{I.93}$$

where we use that ϕ is L-Lipschitz and the definition in (I.82). We also have

$$\begin{aligned}\mathbb{E}_V\left[H_k\right] &= \mathbb{E}_V[\bar{K}]^{-1/2}\mathbb{E}_V\left[\bar{\phi}(Xv_k)\bar{\phi}(Xv_k^\top)\right]\mathbb{E}_V[\bar{K}]^{-1/2} \\ &= \frac{1}{p}\mathbb{E}_V[\bar{K}]^{-1/2}\mathbb{E}_V[\bar{K}]\mathbb{E}_V[\bar{K}]^{-1/2} \\ &= \frac{1}{p}I,\end{aligned} \tag{I.94}$$

which allows us to write

$$\begin{aligned}\mathbb{E}_V\left[(H_k-\mathbb{E}_V[H_k])^2\right] &= \mathbb{E}_V\left[H_k^2\right]-\mathbb{E}_V\left[H_k\right]^2\\ &\preceq \mathbb{E}_V\left[H_k^2\right]\\ &\preceq \mathbb{E}_V\left[\sup_{v_k}\|H_k\|_{\mathsf{op}}H_k\right]\\ &\preceq C\frac{n\log^2 p}{p}\mathbb{E}_V\left[H_k\right]\\ &= C\frac{n\log^2 p}{p^2}I,\end{aligned} \tag{I.95}$$

where the third line holds since H_k is p.s.d. for all v_k, the fourth line is a consequence of (I.92), and the last step of (I.94). This readily gives

$$\left\|\mathbb{E}_V\left[(H_k-\mathbb{E}_V[H_k])^2\right]\right\|_{\mathsf{op}}\leq\frac{Cn\log^2 p}{p^2}. \tag{I.96}$$

Thus, (I.91) is the sum of p independent (in the probability space of V), mean-0, $n\times n$ symmetric random matrices, such that $\|H_k-\mathbb{E}_V[H_k]\|_{\mathsf{op}}\leq\frac{Cn\log^2 p}{p}$ almost surely for all k. Furthermore, we also have $\left\|\mathbb{E}_V\left[(H_k-\mathbb{E}_V[H_k])^2\right]\right\|_{\mathsf{op}}\leq\frac{Cn\log^2 p}{p^2}$. Then, by Theorem 5.4.1 Vershynin [2018], we get

$$\mathbb{P}_V\left(\left\|\mathbb{E}_V[\bar{K}]^{-1/2}\left(\bar{K}-\mathbb{E}_V[\bar{K}]\right)\mathbb{E}_V[\bar{K}]^{-1/2}\right\|_{\mathsf{op}}\geq t\right)\leq 2n\exp\left(-\frac{t^2/2}{\frac{Cn\log^2 p}{p}+\frac{Cn\log^2 p}{3p}t}\right). \tag{I.97}$$

Setting

$$t=\sqrt{\frac{n}{p}}\log p\log n\leq C_4, \tag{I.98}$$

where the last step holds since $n=O\left(p/\log^4 p\right)$, we get

$$\begin{aligned}\mathbb{P}_V\left(\left\|\mathbb{E}_V[\bar{K}]^{-1/2}\left(\bar{K}-\mathbb{E}_V[\bar{K}]\right)\mathbb{E}_V[\bar{K}]^{-1/2}\right\|_{\mathsf{op}}\geq\sqrt{\frac{n}{p}}\log n\log p\right)&\\ \leq 2\exp\left(\log n-\frac{\log^2 n}{2C+2Ct/3}\right)&\\ \leq 2\exp\left(\log n-\frac{\log^2 n}{2C+2CC_4/3}\right)&\\ \leq 2\exp\left(-c_6\log^2 n\right),&\end{aligned} \tag{I.99}$$

which gives the result in (I.81) on the truncated kernel $\bar{K}$. We will now translate this in the desired result. Note that, by (I.86), we have

$$\left\|\mathbb{E}_V[K]^{-1/2}\left(K-\mathbb{E}_V[K]\right)\mathbb{E}_V[K]^{-1/2}\right\|_{\mathsf{op}}=\left\|\mathbb{E}_V[K]^{-1/2}\left(\bar{K}-\mathbb{E}_V[K]\right)\mathbb{E}_V[K]^{-1/2}\right\|_{\mathsf{op}}, \tag{I.100}$$

with probability at least $1-2\exp\left(-c_2\log^2 n\right)$ over V. Applying the triangle inequality we get

$$\begin{aligned}
&\left\|\mathbb{E}_V[K]^{-1/2}\left(\bar{K}-\mathbb{E}_V[K]\right)\mathbb{E}_V[K]^{-1/2}\right\|_{\mathsf{op}}\\
&\leq \left\|\mathbb{E}_V[K]^{-1/2}\left(\bar{K}-\mathbb{E}_V[\bar{K}]\right)\mathbb{E}_V[K]^{-1/2}\right\|_{\mathsf{op}}+\left\|\mathbb{E}_V[K]^{-1/2}\left(\mathbb{E}_V[\bar{K}]-\mathbb{E}_V[K]\right)\mathbb{E}_V[K]^{-1/2}\right\|_{\mathsf{op}}\\
&\leq \left\|\mathbb{E}_V[K]^{-1/2}\left(\bar{K}-\mathbb{E}_V[\bar{K}]\right)\mathbb{E}_V[K]^{-1/2}\right\|_{\mathsf{op}}+\left\|\mathbb{E}_V[K]^{-1/2}\right\|_{\mathsf{op}}^2\left\|\mathbb{E}_V[\bar{K}]-\mathbb{E}_V[K]\right\|_{\mathsf{op}}\\
&\leq \left\|\mathbb{E}_V[K]^{-1/2}\left(\bar{K}-\mathbb{E}_V[\bar{K}]\right)\mathbb{E}_V[K]^{-1/2}\right\|_{\mathsf{op}}+\frac{C_5}{p}\exp\left(-c_4\log^2 p\right),
\end{aligned} \tag{I.101}$$

where the last step follows from (I.88) and Lemma I.9, and it holds with probability at least $1-2\exp\left(-c_7\log^2 n\right)$ over X. Then, we have

$$\begin{aligned}
&\left\|\mathbb{E}_V[K]^{-1/2}\left(\bar{K}-\mathbb{E}_V[\bar{K}]\right)\mathbb{E}_V[K]^{-1/2}\right\|_{\mathsf{op}}\\
&\left\|\mathbb{E}_V[K]^{-1/2}\mathbb{E}_V[\bar{K}]^{1/2}\mathbb{E}_V[\bar{K}]^{-1/2}\left(\bar{K}-\mathbb{E}_V[\bar{K}]\right)\mathbb{E}_V[\bar{K}]^{-1/2}\mathbb{E}_V[\bar{K}]^{1/2}\mathbb{E}_V[K]^{-1/2}\right\|_{\mathsf{op}}\\
&\leq \left\|\mathbb{E}_V[K]^{-1/2}\mathbb{E}_V[\bar{K}]^{1/2}\right\|_{\mathsf{op}}^2\left\|\mathbb{E}_V[\bar{K}]^{-1/2}\left(\bar{K}-\mathbb{E}_V[\bar{K}]\right)\mathbb{E}_V[\bar{K}]^{-1/2}\right\|_{\mathsf{op}}\\
&\leq 2\left\|\mathbb{E}_V[\bar{K}]^{-1/2}\left(\bar{K}-\mathbb{E}_V[\bar{K}]\right)\mathbb{E}_V[\bar{K}]^{-1/2}\right\|_{\mathsf{op}},
\end{aligned} \tag{I.102}$$

where the last step holds since

$$\begin{aligned}
\left\|\mathbb{E}_V[K]^{-1/2}\mathbb{E}_V[\bar{K}]^{1/2}\right\|_{\mathsf{op}}^2 &= \left\|\mathbb{E}_V[K]^{-1/2}\mathbb{E}_V[\bar{K}]\mathbb{E}_V[K]^{-1/2}\right\|_{\mathsf{op}}\\
&\leq \|I\|_{\mathsf{op}}+\left\|\mathbb{E}_V[K]^{-1/2}\left(\mathbb{E}_V[\bar{K}]-\mathbb{E}_V[K]\right)\mathbb{E}_V[K]^{-1/2}\right\|_{\mathsf{op}}\\
&\leq 1+\frac{C_5}{p}\exp\left(-c_4\log^2 p\right),
\end{aligned} \tag{I.103}$$

and the last step of (I.103) follows from (I.88) and Lemma I.9, and it holds with probability at least $1-2\exp\left(-c_7\log^2 n\right)$ over X. Thus, using consecutively (I.100), (I.101), and (I.102), we get

$$\begin{aligned}
&\left\|\mathbb{E}_V[K]^{-1/2}\left(K-\mathbb{E}_V[K]\right)\mathbb{E}_V[K]^{-1/2}\right\|_{\mathsf{op}}\\
&=\left\|\mathbb{E}_V[K]^{-1/2}\left(\bar{K}-\mathbb{E}_V[K]\right)\mathbb{E}_V[K]^{-1/2}\right\|_{\mathsf{op}}\\
&\leq \left\|\mathbb{E}_V[K]^{-1/2}\left(\bar{K}-\mathbb{E}_V[\bar{K}]\right)\mathbb{E}_V[K]^{-1/2}\right\|_{\mathsf{op}}+\frac{C_5}{p}\exp\left(-c_4\log^2 p\right)\\
&\leq 2\left\|\mathbb{E}_V[\bar{K}]^{-1/2}\left(\bar{K}-\mathbb{E}_V[\bar{K}]\right)\mathbb{E}_V[\bar{K}]^{-1/2}\right\|_{\mathsf{op}}+\frac{C_5}{p}\exp\left(-c_4\log^2 p\right)\\
&=O\left(\sqrt{\frac{n}{p}}\log n\log p\right),
\end{aligned} \tag{I.104}$$

where the last step follows from (I.99), and the steps jointly hold with probability at least $1-2\exp\left(-c_8\log^2 n\right)$ over X and V. This concludes the proof. □

Lemma I.11. *Let Assumptions 8.5 and 8.6 hold, and let* $n=o\left(p/\log^4 p\right)$, $n\log^3 n=o\left(d^{3/2}\right)$ *and* $n=\omega(d)$. *Then, with probability at least* $1-2\exp\left(-c\log^2 n\right)$ *over* X *and* V, *all the following hold*

$$\lambda_d(K)=\Omega\left(\frac{pn}{d}\right), \tag{I.105}$$

$$\lambda_{d+1}(K) = O\left(p\right), \tag{I.106}$$

$$\lambda_n(K) = \lambda_{\min}\left(K\right) = \Omega(p), \tag{I.107}$$

where c is an absolute constant.

Proof. By Lemma I.10, we have that, with probability at least $1 - 2\exp\left(-c_1 \log^2 n\right)$ over X and V, for all $u \in \mathbb{R}^n$ we have

$$u^\top \left(\mathbb{E}_V[K]^{-1/2}\left(K - \mathbb{E}_V[K]\right)\mathbb{E}_V[K]^{-1/2}\right) u \leq C\sqrt{\frac{n}{p}} \log n \log p \left\|u\right\|_2^2, \tag{I.108}$$

where C is an absolute constant. Then, setting $u = \mathbb{E}_V[K]^{1/2}\hat{u}$, we have that, for all $\hat{u} \in \mathbb{R}^n$,

$$\hat{u}^\top \left(K - \mathbb{E}_V[K]\right)\hat{u} \preceq \hat{u}^\top \left(C\sqrt{\frac{n}{p}} \log n \log p\, \mathbb{E}_V[K]\right)\hat{u}, \tag{I.109}$$

which reads

$$K - \mathbb{E}_V[K] \preceq C\sqrt{\frac{n}{p}} \log n \log p\, \mathbb{E}_V[K]. \tag{I.110}$$

In the same way, considering instead $(\mathbb{E}_V[K] - K)$ in the very first equation, we can also derive

$$\mathbb{E}_V[K] - K \preceq C\sqrt{\frac{n}{p}} \log n \log p\, \mathbb{E}_V[K], \tag{I.111}$$

which therefore gives

$$\left(1 - C\sqrt{\frac{n}{p}} \log n \log p\right)\mathbb{E}_V[K] \preceq K \preceq \left(1 + C\sqrt{\frac{n}{p}} \log n \log p\right)\mathbb{E}_V[K]. \tag{I.112}$$

By the Courant–Fischer–Weyl min-max principle, we can write

$$\lambda_i(K) = \max_S \min_{u \in S,\, \left\|u\right\|_2 = 1} \left(u^\top K u\right), \tag{I.113}$$

where S is any i-dimensional subspace of $\mathbb{R}^n$. Let $S^* = \arg\max_S \min_{u \in S,\, \left\|u\right\|_2 = 1} \left(u^\top K u\right)$. Then, by (I.112), we have

$$\begin{aligned}
\lambda_i(K) &= \min_{u \in S^*,\, \left\|u\right\|_2 = 1} \left(u^\top K u\right) \\
&\leq \min_{u \in S^*,\, \left\|u\right\|_2 = 1} \left(u^\top \left(\left(1 + C\sqrt{\frac{n}{p}} \log n \log p\right)\mathbb{E}_V[K]\right) u\right) \\
&\leq \max_S \min_{u \in S,\, \left\|u\right\|_2 = 1} \left(u^\top \left(\left(1 + C\sqrt{\frac{n}{p}} \log n \log p\right)\mathbb{E}_V[K]\right) u\right) \\
&= \left(1 + C\sqrt{\frac{n}{p}} \log n \log p\right) \lambda_i\left(\mathbb{E}_V[K]\right),
\end{aligned} \tag{I.114}$$

where in the last line we use the same principle to express the i-th eigenvalue of $\mathbb{E}_V[K]$. Following the same strategy, we can therefore conclude that

$$\left(1 - C\sqrt{\frac{n}{p}} \log n \log p\right) \lambda_i\left(\mathbb{E}_V[K]\right) \leq \lambda_i\left(K\right) \leq \left(1 + C\sqrt{\frac{n}{p}} \log n \log p\right) \lambda_i\left(\mathbb{E}_V[K]\right). \tag{I.115}$$

By Lemma I.9, we have that, with probability at least $1-2\exp\left(-c_2\log n\right)$ over X, $\lambda_d\left(\mathbb{E}_V[K]\right) = \Omega(np/d)$, and that $\lambda_i\left(\mathbb{E}_V[K]\right) = \Theta(p)$, for $d+1 \leq i \leq n$. This implies

$$\lambda_d(K) = \Omega\left(\left(1 - C\sqrt{\frac{n}{p}}\log n \log p\right)\frac{np}{d}\right) = \Omega\left(\frac{pn}{d}\right), \tag{I.116}$$

$$\lambda_{d+1}(K) = O\left(\left(1 + C\sqrt{\frac{n}{p}}\log n \log p\right)p\right) = O\left(p\right), \tag{I.117}$$

$$\lambda_n(K) = \lambda_{\min}\left(K\right) = \Omega\left(\left(1 - C\sqrt{\frac{n}{p}}\log n \log p\right)p\right) = \Omega(p), \tag{I.118}$$

where we use $n\log^2 n = o(p/\log^2 p)$. This gives the desired result. □

Lemma I.12. *Let Assumptions 8.5 and 8.6 hold, and let $d = o(p)$ and $d = o(n)$. Then, we have*

$$\|\Phi\|_{\mathsf{op}} = O\left(\sqrt{\frac{np}{d}}\right), \tag{I.119}$$

with probability at least $1 - 2\exp\left(-cd\right)$ over X and V, where c is an absolute constant.

Proof. The k-th row of $\Phi^\top$ takes the form $\phi(Xv_k) \in \mathbb{R}^n$. Since the Gaussian distribution is Lipschitz concentrated (see Theorem 5.22 in Vershynin [2018]), we have that

$$\left\|\phi(Xv_k)\right\|_{\psi_2} = \left\|\phi(Xv_k) - \mathbb{E}_{v_k}\left[\phi(Xv_k)\right]\right\|_{\psi_2} \leq C_1\frac{\|X\|_{\mathsf{op}}}{\sqrt{d}}, \tag{I.120}$$

where the first step holds since the 0-th Hermite coefficient of ϕ is zero, and the second step holds since ϕ is Lipschitz. Notice that the term $\sqrt{d}$ is due to the fact that $\sqrt{d}\,v_k$ is standard Gaussian. We remark that the sub-Gaussian norm in this equation is intended in the probability space of v_k. By Lemma I.8, we have that

$$\|X\|_{\mathsf{op}} = O\left(\sqrt{n}\right), \tag{I.121}$$

with probability at least $1 - 2\exp\left(-c_1 d\right)$ over X. Thus, conditioning on this high probability event, due to (I.120), we have that $\Phi^\top$ is a $p \times n$ matrix whose rows are i.i.d. mean-0 random vectors with sub-Gaussian norm $O\left(\sqrt{n/d}\right)$. Then, by Lemma B.7 of Bombari et al. [2022b], we have

$$\left\|\Phi^\top\right\|_{\mathsf{op}} = O\left(\sqrt{\frac{np}{d}}\right), \tag{I.122}$$

with probability at least $1 - 2\exp(-c_2 n)$ over V, which provides the desired result. □

Lemma I.13. *Let Assumptions 8.5 and 8.6 hold, and let $n\log^3 n = O\left(d^{3/2}\right)$. Then, we have*

$$\left\|\mathbb{E}_V\left[\tilde{K}\right]\right\|_{\mathsf{op}} = O\left(p\right), \tag{I.123}$$

with probability at least $1 - 2\exp\left(-c\log^2 n\right)$ over X, where c is an absolute constant.

Proof. The proof follows the same strategy as the proof of Lemma I.9, with the difference that now there is no term M_1. In fact, $\tilde{\phi}$ has the same Hermite coefficients as ϕ, except that $\tilde{\mu}_1 = 0$. The thesis then follows from (I.76). We remark that to prove this result, differently from the proof of Lemma I.9, $n = \omega(d)$ is not required. Also, we require only $n\log^3 n = O\left(d^{3/2}\right)$ instead of $n\log^3 n = o\left(d^{3/2}\right)$. In fact, we do not need a bound on the smallest eigenvalue of M_2 (see (I.76)), and therefore we just need $\left\|M_2 - \tilde{M}_2\right\|_{\mathsf{op}} = O\left(p\right)$ (see (I.75)). □

Lemma I.14. *Let Assumptions 8.5 and 8.6 hold, and let* $n = O\left(p/\log^4 p\right)$ *and* $n\log^3 n = O\left(d^{3/2}\right)$. *Then, we have*

$$\left\|\tilde{\Phi}\right\|_{\mathsf{op}} = O\left(\sqrt{p}\right), \tag{I.124}$$

with probability at least $1 - 2\exp\left(-c\log^2 n\right)$ *over* X *and* V, *where* c *is an absolute constant.*

Proof. Consider the truncated function

$$\bar{\phi}(z) := \tilde{\phi}(z)\mathbf{1}\left(|z| \leq \frac{\log p}{\tilde{L}}\right), \tag{I.125}$$

where $\tilde{L}$ is the Lipschitz constant of $\tilde{\phi}$. Define the truncated kernel $\bar{K}$ accordingly

$$\bar{K}_{ij} := \bar{\phi}(Vx_i)^\top \bar{\phi}(Vx_j). \tag{I.126}$$

We now compare $\mathbb{E}_V\left[\bar{K}\right]$ and $\mathbb{E}_V[\tilde{K}]$. Let $v \sim \mathcal{N}(0, I/d)$ and define the random variable

$$E_{ij} := \mathbf{1}\left(|v^\top x_i| > \frac{\log p}{\tilde{L}} \text{ or } |v^\top x_j| > \frac{\log p}{\tilde{L}}\right). \tag{I.127}$$

Note that since E_{ij} is an indicator we have $E_{ij} = E_{ij}^2$, and

$$\mathbb{P}_V(E_{ij} = 1) \leq \mathbb{P}_V\left(|v^\top x_i| > \frac{\log p}{\tilde{L}}\right) + \mathbb{P}_V\left(|v^\top x_j| > \frac{\log p}{\tilde{L}}\right) \leq \exp\left(-c_1\log^2 p\right), \tag{I.128}$$

where the second step holds as $v^\top x_i \sim \mathcal{N}(0,1)$ in the probability space of V.

Thus, we can write

$$\begin{aligned}
\left|\mathbb{E}_V\left[\bar{K}_{ij}\right] - \mathbb{E}_V[\tilde{K}_{ij}]\right| &= p\left|\mathbb{E}_v\left[\tilde{\phi}(v^\top x_i)\tilde{\phi}(v^\top x_j)\, E_{ij}\right]\right| \\
&\leq p\,\mathbb{E}_v\left[\left(\tilde{\phi}(v^\top x_i)\tilde{\phi}(v^\top x_j)\right)^2\right]^{1/2} \mathbb{E}_v\left[E_{ij}^2\right]^{1/2} \\
&\leq p\,\mathbb{E}_v\left[\left(\tilde{\phi}(v^\top x_i)\right)^4\right]^{1/2} \mathbb{P}_V(E_{ij} = 1)^{1/2} \\
&\leq p\,C^{1/2}\exp\left(-\frac{c_1\log^2 p}{2}\right) \\
&\leq p\exp\left(-c_2\log^2 p\right).
\end{aligned} \tag{I.129}$$

Here, the second and third lines follow from Cauchy-Schwartz inequality; the fourth line is a consequence of (I.128) and $\tilde{\phi}(v^\top x_i)$ being a sub-Gaussian random variable ($\tilde{\phi}(z)$ is Lipschitz), and thus with bounded fourth moment (see Equation 2.11 in Vershynin [2018]). This result holds for any i, j, and therefore implies

$$\begin{aligned}
\left\|\mathbb{E}_V\left[\bar{K}\right] - \mathbb{E}_V[\tilde{K}]\right\|_{\mathsf{op}} &\leq \left\|\mathbb{E}_V\left[\bar{K}\right] - \mathbb{E}_V[\tilde{K}]\right\|_F \\
&\leq pn\exp\left(-c_2\log^2 p\right) \\
&\leq \exp\left(-c_3\log^2 p\right),
\end{aligned} \tag{I.130}$$

where in the last step we used $n = O(p)$, which follows from Assumption 8.7. This last equation, together with Lemma I.13, gives

$$\left\|\mathbb{E}_V\left[\bar{K}\right]\right\|_{\mathsf{op}} = O(p), \tag{I.131}$$

with probability at least $1 - \exp\left(-c_4 \log^2 n\right)$. We will condition on this high probability event until the end of the proof.

We now define the matrix $\bar{H}_k \in \mathbb{R}^{n\times n}$ as

$$\bar{H}_k = \bar{\phi}(Xv_k)\bar{\phi}(Xv_k)^\top, \tag{I.132}$$

which implies

$$\bar{\Phi}\bar{\Phi}^\top - \mathbb{E}_V\left[\bar{\Phi}\bar{\Phi}^\top\right] = \sum_{k=1}^{p} \bar{H}_k - \mathbb{E}_V\left[\bar{H}_k\right]. \tag{I.133}$$

Now, we have

$$\sup_{v_k}\left\|\bar{H}_k - \mathbb{E}_V[\bar{H}_k]\right\|_{\mathsf{op}} \le 2\sup_{v_k}\left\|\bar{H}_k\right\|_{\mathsf{op}} \le 2\sup_{v_k}\left\|\bar{\phi}(Xv_k)\right\|_2^2 \le 2n\log^2 p, \tag{I.134}$$

where the first step is true because of Jensen and triangle inequality, and the last step holds because, for every $i \in [n]$,

$$\left|\bar{\phi}(v_k^\top x_i)\right| \le \tilde{L}\frac{\log p}{\tilde{L}} = \log p, \tag{I.135}$$

given the definition in (I.125). Also, note that

$$\mathbb{E}_V\left[\bar{H}_k\right] = \mathbb{E}_V\left[\bar{\phi}(Xv_k)\bar{\phi}(Xv_k)^\top\right] = \frac{1}{p}\mathbb{E}_V[\bar{K}], \tag{I.136}$$

which gives, together with (I.131),

$$\left\|\mathbb{E}_V\left[\bar{H}_k\right]\right\|_{\mathsf{op}} = O(1). \tag{I.137}$$

Then, we can use (I.134) and (I.137), together with the same argument in (I.95), to get

$$\left\|\mathbb{E}_V\left[\left(\bar{H}_k - \mathbb{E}_V[\bar{H}_k]\right)^2\right]\right\|_{\mathsf{op}} = O\left(n\log^2 p\right). \tag{I.138}$$

Thus, (I.133) is the sum of p independent (in the probability space of V), mean-0, $n \times n$ symmetric random matrices, such that $\left\|\bar{H}_k - \mathbb{E}_V[\bar{H}_k]\right\|_{\mathsf{op}} \le 2n\log^2 p$ almost surely for all k, and $\left\|\mathbb{E}_V\left[\left(\bar{H}_k - \mathbb{E}_V[\bar{H}_k]\right)^2\right]\right\|_{\mathsf{op}} \le Cn\log^2 p$, for some constant C. Then, by Theorem 5.4.1 of [Vershynin, 2018], we get

$$\mathbb{P}\left(\left\|\bar{\Phi}\bar{\Phi}^\top - \mathbb{E}_V\left[\bar{\Phi}\bar{\Phi}^\top\right]\right\|_{\mathsf{op}} \ge t\right) \le 2n\exp\left(-\frac{t^2/2}{Cnp\log^2 p + 2tn\log^2 p/3}\right). \tag{I.139}$$

Setting $t = p$, we obtain

$$\mathbb{P}\left(\left\|\bar{\Phi}\bar{\Phi}^\top - \mathbb{E}_V\left[\bar{\Phi}\bar{\Phi}^\top\right]\right\|_{\mathsf{op}} \ge p\right) \le 2\exp\left(-c_5\frac{p}{n\log^2 p}\right) \le 2\exp\left(-c_6\log^2 n\right), \tag{I.140}$$

where in the last step we used $n = O\left(p/\log^4 p\right)$. To conclude, we have that

$$\begin{aligned}\mathbb{P}_V(\tilde{\Phi}\tilde{\Phi}^\top \neq \bar{\Phi}\bar{\Phi}^\top) &\leq \mathbb{P}_V\left(\max_{i\in[n],k\in[p]} |v_k^\top x_i| > \log p/\tilde{L}\right) \\ &\leq 2np\exp\left(-c_6\log^2 p\right) \\ &\leq 2\exp\left(-c_7\log^2 p\right),\end{aligned} \tag{I.141}$$

where the second step holds as $v_k^\top x_i \sim \mathcal{N}(0,1)$ in the probability space of V, and the last step holds because of the assumption $n = O\left(p/\log^4 p\right)$. Putting (I.140) and (I.141) together, we finally get

$$\left\|\tilde{\Phi}\tilde{\Phi}^\top\right\|_{\mathsf{op}} = \left\|\bar{\Phi}\bar{\Phi}^\top\right\|_{\mathsf{op}} \leq \left\|\mathbb{E}_V\left[\bar{\Phi}\bar{\Phi}^\top\right]\right\|_{\mathsf{op}} + \left\|\bar{\Phi}\bar{\Phi}^\top - \mathbb{E}_V\left[\bar{\Phi}\bar{\Phi}^\top\right]\right\|_{\mathsf{op}} = O\left(p\right), \tag{I.142}$$

with probability at least $1 - 2\exp\left(-c\log^2 n\right)$ over X and V, which gives the thesis. □

Lemma I.15. *Let Assumptions 8.5 and 8.6 hold, and $p = \omega(n)$. Then, we have*

$$\left|\|\varphi(x_1)\|_2^2 - Mp\right| = O\left(\sqrt{p}\log n\right), \tag{I.143}$$

$$\left|\|\tilde{\varphi}(x_1)\|_2^2 - M_1 p\right| = O\left(\sqrt{p}\log n\right), \tag{I.144}$$

with M and M_1 being two positive constants only depending on ϕ, with probability at least $1 - 2\exp(-c\log^2 n)$ over V, where c is an absolute constant.

Furthermore, we have that $M - M_1 = \mu_1^2 > 0$, which, conditioning on the previous result, implies

$$\|\varphi(x_1)\|_2 - \|\tilde{\varphi}(x_1)\|_2 = \Omega(\sqrt{p}). \tag{I.145}$$

Proof. We have

$$\|\varphi(x_1)\|_2^2 = \sum_{k=1}^{p} \phi(v_k^\top x_1)^2, \tag{I.146}$$

where we use the shorthand v_k to indicate the k-th row of V. As ϕ is Lipschitz, $v_j \sim \mathcal{N}(0, I/d)$ and $\|x_1\|_2 = \sqrt{d}$, we have that $\|\varphi(x_1)\|_2^2$ is the sum of p independent sub-exponential random variables, in the probability space of V. Thus, by Bernstein inequality (see Theorem 2.8.1 in Vershynin [2018]), we have

$$\left|\|\varphi(x_1)\|_2^2 - \mathbb{E}_V\left[\|\varphi(x_1)\|_2^2\right]\right| = O\left(\sqrt{p}\log n\right), \tag{I.147}$$

with probability at least $1 - \exp(-c_1\log^2 n)$, over the probability space of V. Exploiting the Hermite expansion of ϕ, we get

$$\mathbb{E}_V\left[\|\varphi(x_1)\|_2^2\right] = p\,\mathbb{E}_{\rho\sim(0,1)}\left[\phi(\rho)^2\right] = pM, \tag{I.148}$$

where we set $M = \sum_{l=0}^{\infty}\mu_l^2$.

The same argument applied on $\tilde{\phi}$, which is also Lipschitz, guarantees

$$\left|\|\tilde{\varphi}(x_1)\|_2^2 - M_1 p\right| = O\left(\sqrt{p}\log n\right), \tag{I.149}$$

where $M_1 = \sum_{l=3}^{\infty}\mu_l^2$, since the 0-th, 1st and 2nd Hermite coefficients of $\tilde{\phi}$ are 0. Thus, we readily get

$$M - M_1 = \mu_1^2. \tag{I.150}$$

To conclude, it is sufficient to notice that, conditioning on the previous two high probability events, we have that $\|\varphi(x_1)\|_2$ and $\|\tilde{\varphi}(x_1)\|_2$ are both $O\left(\sqrt{p}\right)$. Then,

$$\begin{aligned}\|\varphi(x_1)\|_2 - \|\tilde{\varphi}(x_1)\|_2 &= \frac{\|\varphi(x_1)\|_2^2 - \|\tilde{\varphi}(x_1)\|_2^2}{\|\varphi(x_1)\|_2 + \|\tilde{\varphi}(x_1)\|_2} \\ &\geq \frac{(M - M_1)p - \left|\|\varphi(x_1)\|_2^2 - pM\right| - \left|\|\tilde{\varphi}(x_1)\|_2^2 - M_1 p\right|}{\|\varphi(x_1)\|_2 + \|\tilde{\varphi}(x_1)\|_2} \\ &= \Omega(\sqrt{p}),\end{aligned} \tag{I.151}$$

where we use the triangle inequality twice in the second line and conclude using $\sqrt{p} = \omega(\log n)$. □

Lemma I.16. *Let Assumptions 8.5 and 8.6 hold, and let $n = o\left(p/\log^4 p\right)$, $n\log^3 n = o\left(d^{3/2}\right)$ and $n = \omega(d)$. Then, we jointly have*

$$\left\|X^+\right\|_{\mathsf{op}} = O\left(\frac{1}{\sqrt{n}}\right), \tag{I.152}$$

$$\left\|V^+\right\|_{\mathsf{op}} = O\left(\sqrt{\frac{d}{p}}\right), \tag{I.153}$$

$$\left\|\mu_1 V^\top - X^+\Phi\right\|_{\mathsf{op}} = O\left(\sqrt{\frac{p}{n}}\right), \tag{I.154}$$

$$\left\|\mu_1 V^\top \Phi^+\right\|_{\mathsf{op}} = O\left(\frac{1}{\sqrt{n}}\right), \tag{I.155}$$

with probability at least $1 - 2\exp\left(-c\log^2 n\right)$ over X and V, where c is an absolute constant.

Proof. The first two statements easily follow from Lemma I.8, which gives

$$\left\|X^+\right\|_{\mathsf{op}} = \lambda_{\min}\left(X^\top X\right)^{-1/2} = O\left(\frac{1}{\sqrt{n}}\right), \tag{I.156}$$

$$\left\|V^+\right\|_{\mathsf{op}} = \lambda_{\min}\left(V^\top V\right)^{-1/2} = O\left(\sqrt{\frac{d}{p}}\right), \tag{I.157}$$

with probability at least $1 - 2\exp\left(-c_1 d\right) \geq 1 - 2\exp\left(-c_2\log^2 n\right)$ over X and V. For the third statement, we condition on this high probability event, and notice that

$$\Phi = \mu_1 X V^\top + \tilde{\Phi}, \tag{I.158}$$

which gives

$$\left\|\mu_1 V^\top - X^+\Phi\right\|_{\mathsf{op}} = \left\|X^+\tilde{\Phi}\right\|_{\mathsf{op}} \leq \left\|X^+\right\|_{\mathsf{op}}\left\|\tilde{\Phi}\right\|_{\mathsf{op}} = O\left(\sqrt{\frac{p}{n}}\right), \tag{I.159}$$

where the last step holds because of Lemma I.14 with probability at least $1 - 2\exp\left(-c_3\log^2 n\right)$. For the fourth statement, we write

$$\begin{aligned}\left\|\mu_1 V^\top \Phi^+\right\|_{\mathsf{op}} &\leq \left\|\left(\mu_1 V^\top - X^+\Phi\right)\Phi^+\right\|_{\mathsf{op}} + \left\|X^+\Phi\Phi^+\right\|_{\mathsf{op}} \\ &\leq \left\|\mu_1 V^\top - X^+\Phi\right\|_{\mathsf{op}}\left\|\Phi^+\right\|_{\mathsf{op}} + \left\|X^+\right\|_{\mathsf{op}} \\ &= O\left(\sqrt{\frac{p}{n}}\sqrt{\frac{1}{p}} + \frac{1}{\sqrt{n}}\right) = O\left(\frac{1}{\sqrt{n}}\right),\end{aligned} \tag{I.160}$$

where the second step follows from

$$\left\|\Phi^{+}\right\|_{\mathsf{op}} = \lambda_{\min}\left(K\right)^{-1/2} = O\left(\frac{1}{\sqrt{p}}\right), \tag{I.161}$$

which holds with probability $1 - 2\exp\left(-c_4 \log^2 n\right)$ because of Lemma I.11. This final result provides the thesis. □

Lemma I.17. *Let Assumptions 8.5 and 8.6 hold, and let $n = o\left(p/\log^4 p\right)$, $n\log^3 n = o\left(d^{3/2}\right)$ and $n = \omega(d)$. Let $P_\Lambda \in \mathbb{R}^{p\times p}$ be the projector on the span of the d eigenvectors corresponding to the d highest eigenvalues of $\Phi^\top\Phi$, and P_V be the projector on the span of the columns of V. Then, we have that*

$$\left\|P_\Lambda^\perp V\right\|_{\mathsf{op}} = O\left(\sqrt{\frac{p}{n}}\right), \tag{I.162}$$

and

$$\left\|P_\Lambda^\perp P_V\right\|_{\mathsf{op}} = O\left(\sqrt{\frac{d}{n}}\right), \tag{I.163}$$

jointly hold with probability at least $1 - 2\exp\left(-c\log^2 n\right)$ over X and V, where c is an absolute constant.

Proof. Let $u \in \mathbb{R}^p$ be a vector such that $\left\|u\right\|_2 = 1$. Then, we have

$$\left\|\Phi P_\Lambda^\perp u\right\|_2 \le \sqrt{\lambda_{d+1}(\Phi^\top\Phi)} = \sqrt{\lambda_{d+1}(K)} = O\left(\sqrt{p}\right), \tag{I.164}$$

where the first inequality follows by the definition of P_Λ, and $\lambda_{d+1}(K)$ ($\lambda_{d+1}(\Phi^\top\Phi)$) is the $(d+1)$-th eigenvalue of K ($\Phi^\top\Phi$) when sorting them in non-increasing order. Then, the last equality holds with probability at least $1 - 2\exp\left(-c_1\log^2 n\right)$ over X and V, because of Lemma I.11. Then, conditioning on the previous high probability event, since $\Phi = \tilde{\Phi} + \mu_1 X V^\top$, we can write

$$\begin{aligned}
\mu_1\sqrt{\lambda_{\min}\left(X^\top X\right)}\left\|V^\top P_\Lambda^\perp u\right\|_2 &\le \left\|\mu_1 X V^\top P_\Lambda^\perp u\right\|_2 \\
&\le \left\|\tilde{\Phi} P_\Lambda^\perp u\right\|_2 + \left\|\Phi P_\Lambda^\perp u\right\|_2 \\
&\le \left\|\tilde{\Phi}\right\|_{\mathsf{op}}\left\|P_\Lambda^\perp\right\|_{\mathsf{op}}\left\|u\right\|_2 + \left\|\Phi P_\Lambda^\perp u\right\|_2 = O\left(\sqrt{p}\right),
\end{aligned} \tag{I.165}$$

where the last step is a consequence of Lemma I.14, and it holds with probability $1 - 2\exp\left(-c_2\log^2 n\right)$. By Lemma I.8, we have that $\lambda_{\min}\left(X^\top X\right) = \Theta(n)$ with probability at least $1 - 2\exp\left(-c_3 d\right)$. Conditioning on such high probability events, (I.165) gives

$$\left\|V^\top P_\Lambda^\perp u\right\|_2 = O\left(\sqrt{\frac{p}{n}}\right), \tag{I.166}$$

which provides the first part of the thesis, since it uniformly holds for every u (as we did not condition on any event dependent on u itself). To conclude, since we have

$$P_V = (V^+)^\top V^\top, \tag{I.167}$$

we can use Lemma I.16 to write

$$\left\|P_V P_\Lambda^\perp u\right\|_2 = \left\|(V^+)^\top V^\top P_\Lambda^\perp u\right\|_2 \le \left\|V^+\right\|_{\mathsf{op}}\left\|V^\top P_\Lambda^\perp u\right\|_2 = O\left(\sqrt{\frac{d}{n}}\right), \tag{I.168}$$

with probability at least $1 - 2\exp\left(-c_4 d\right)$ over V. Since again this result holds uniformly on all $\left\|u\right\|_2 = 1$ and $\left\|P_\Lambda^\perp P_V\right\|_{\mathsf{op}} = \left\|P_V P_\Lambda^\perp\right\|_{\mathsf{op}}$, the thesis readily follows. □

Lemma I.18. *Let Assumptions 8.5 and 8.6 hold, and let* $n = o\left(p/\log^4 p\right)$, $n\log^3 n = o\left(d^{3/2}\right)$ *and* $n = \omega(d)$. *Let* $P_\Lambda \in \mathbb{R}^{p\times p}$ *be the projector on the span of the* d *eigenvectors corresponding to the* d *highest eigenvalues of* $\Phi^\top\Phi$. *Then, we have*

$$\|P_\Lambda \varphi(x_1)\|_2 = \Omega(\sqrt{p}), \tag{I.169}$$

with probability at least $1 - 2\exp\left(-c\log^2 n\right)$ *over* X *and* V, *where* c *is an absolute constant.*

Proof. Multiple applications of the triangle inequality yield

$$\begin{aligned}
\|P_\Lambda \varphi(x_1)\|_2 &\geq \|\varphi(x_1)\|_2 - \left\|P_\Lambda^\perp \varphi(x_1)\right\|_2 \\
&= \|\varphi(x_1)\|_2 - \left\|P_\Lambda^\perp \left(\mu_1 V x_1 + \tilde{\varphi}(x_1)\right)\right\|_2 \\
&\geq \|\varphi(x_1)\|_2 - \mu_1 \left\|P_\Lambda^\perp V x_1\right\|_2 - \left\|P_\Lambda^\perp \tilde{\varphi}(x_1)\right\|_2 \\
&\geq \|\varphi(x_1)\|_2 - \mu_1 \left\|P_\Lambda^\perp V\right\|_{\mathsf{op}} \|x_1\|_2 - \|\tilde{\varphi}(x_1)\|_2 .
\end{aligned} \tag{I.170}$$

Let us condition on the results of Lemma I.17 and Lemma I.15. Then, we have that

$$\|\varphi(x_1)\|_2 - \|\tilde{\varphi}(x_1)\|_2 = \Omega(\sqrt{p}), \tag{I.171}$$

and

$$\left\|P_\Lambda^\perp V\right\|_{\mathsf{op}} \|x_1\|_2 = O\left(\sqrt{\frac{p}{n}}\right)\sqrt{d} = o(\sqrt{p}), \tag{I.172}$$

which readily gives the thesis. □

Lemma I.19. *We have that*

$$\left\|\tilde{\Phi} V\right\|_{\mathsf{op}} = O\left(\sqrt{\frac{pn}{d}}\log n\right), \tag{I.173}$$

with probability at least $1 - 2\exp\left(-c\log^2 n\right)$ *over* X *and* V, *where* c *is an absolute constant.*

Proof. Note that $\tilde{\phi}$ is Lipschitz (since ϕ is Lipschitz by Assumption 8.6). During the proof, we condition on the event $\|X\|_{\mathsf{op}} = O\left(\sqrt{n}\right)$, which is true with probability at least $1 - 2\exp\left(-c_1 d\right)$ by Lemma I.8, and we use the shorthand $v \in \mathbb{R}^d$ to denote a random vector such that $\sqrt{d}\, v$ is a standard Gaussian vector, *i.e.*, it has the same distribution as the rows of V. This implies

$$\mathbb{E}_v\left[\left\|\tilde{\phi}\left(Xv\right)\right\|_2\right] = O\left(\sqrt{n}\right), \qquad \left\|\left\|\tilde{\phi}\left(Xv\right)\right\|_2 - \mathbb{E}_v\left[\left\|\tilde{\phi}\left(Xv\right)\right\|_2\right]\right\|_{\psi_2} = O\left(\sqrt{\frac{n}{d}}\right), \tag{I.174}$$

and

$$\mathbb{E}_v\left[\|v\|_2\right] = O\left(1\right), \qquad \left\|\|v\|_2 - \mathbb{E}_v\left[\|v\|_2\right]\right\|_{\psi_2} = O\left(\frac{1}{\sqrt{d}}\right), \tag{I.175}$$

where both sub-Gaussian norms are meant on the probability space of v. Here, the very first equation follows from the discussion in Lemma C.3 in Bombari et al. [2022b], and the upper bounds on the sub-Gaussian norms follow from the Lipschitz concentration property of $\sqrt{d}v$. Then, there exists an absolute constant C_1 such that we jointly have

$$\left\|\tilde{\phi}\left(Xv\right)\right\|_2 \leq C_1\sqrt{n}, \qquad \|v\|_2 \leq C_1, \tag{I.176}$$

with probability at least $1 - 2\exp(-c_2 d)$ over v.

Let E_k be the indicator defined on the high probability event above with respect to the random variable $v_k := V_{:k}$ (the k-th row of V), *i.e.*,

$$E_k := \mathbf{1}\left(\|v_k\|_2 \leq C_1 \text{ and } \left\|\tilde{\phi}(Xv_k)\right\|_2 \leq C_1\sqrt{n}\right), \tag{I.177}$$

and we define $E \in \mathbb{R}^{p\times p}$ as the diagonal matrix containing E_k in its k-th entry. Notice that we have $\|I - E\|_{\mathsf{op}} = 0$ with probability at least $1 - 2p\exp(-c_2 d)$, and $\mathbb{E}_V\left[\|I - E\|_{\mathsf{op}}\right] \leq 2p\exp(-c_2 d)$.

Thus, we have

$$\begin{aligned}
\left\|\mathbb{E}_V\left[\tilde{\Phi}(I-E)V\right]\right\|_{\mathsf{op}} &\leq \mathbb{E}_V\left[\left\|\tilde{\Phi}\right\|_{\mathsf{op}}\|1-E\|_{\mathsf{op}}\|V\|_{\mathsf{op}}\right] \\
&\leq \mathbb{E}_V\left[\left\|\tilde{\Phi}\right\|_{\mathsf{op}}^2\|V\|_{\mathsf{op}}^2\right]^{1/2}\mathbb{E}_V\left[\|I-E\|_{\mathsf{op}}^2\right]^{1/2} \\
&\leq \mathbb{E}_V\left[\left\|\tilde{\Phi}\right\|_{\mathsf{op}}^4\right]^{1/4}\mathbb{E}_V\left[\|V\|_{\mathsf{op}}^4\right]^{1/4}(2p\exp(-c_2 d))^{1/2} \\
&\leq \mathbb{E}_V\left[\left\|\tilde{\Phi}\right\|_{F}^4\right]^{1/4}\mathbb{E}_V\left[\|V\|_{F}^4\right]^{1/4}(2p\exp(-c_2 d))^{1/2} \\
&= o(1),
\end{aligned} \tag{I.178}$$

where the last step holds because of our initial conditioning on X: the first two terms are the sum of finite powers of sub-Gaussian random variables (the entries of $\tilde{\Phi}$ and V), and thus (see Proposition 2.5.2 in Vershynin [2018]) the first two factors in the third line of the previous equation will be $O(p^\alpha)$ for some finite α, which gives the last line due to Assumption 8.7.

Exploiting the Hermite expansion of $\tilde{\phi}$ we can write

$$\left[\mathbb{E}_V\left[\tilde{\Phi}V\right]\right]_{ij} = p\left[\mathbb{E}_v\left[\tilde{\phi}(Xv)v^\top\right]\right]_{ij} = \frac{p}{\sqrt{d}}\mathbb{E}_v\left[\tilde{\phi}\left(x_i^\top v\right)\left(e_j^\top\left(\sqrt{d}v\right)\right)\right] = 0, \tag{I.179}$$

where the last step holds since the first Hermite coefficient of $\tilde{\phi}$ is 0. Thus, the application of the triangle inequality to this last equation and (I.178) gives

$$\left\|\mathbb{E}_V\left[\tilde{\Phi}EV\right]\right\|_{\mathsf{op}} \leq \left\|\mathbb{E}_V\left[\tilde{\Phi}V\right]\right\|_{\mathsf{op}} + \left\|\mathbb{E}_V\left[\tilde{\Phi}(I-E)V\right]\right\|_{\mathsf{op}} = o(1), \tag{I.180}$$

with probability at least $1 - 2\exp\left(-c_3\log^2 d\right)$ over X.

Let's now look at

$$\tilde{\Phi}EV - \mathbb{E}_V\left[\tilde{\Phi}EV\right] = \sum_{k=1}^p \tilde{\phi}(Xv_k)E_k v_k^\top - \mathbb{E}_{v_k}\left[\tilde{\phi}(Xv_k)E_k v_k^\top\right] =: \sum_{k=1}^p W_k, \tag{I.181}$$

where we defined the shorthand $W_k = \tilde{\phi}(Xv_k)E_k v_k^\top - \mathbb{E}_{v_k}\left[\tilde{\phi}(Xv_k)E_k v_k^\top\right]$. (I.181) is the sum of p i.i.d. mean-0 random matrices W_k (in the probability space of V), such that

$$\begin{aligned}
\sup_{v_k}\left\|\tilde{\phi}(Xv_k)E_k v_k^\top - \mathbb{E}_{v_k}\left[\tilde{\phi}(Xv_k)E_k v_k^\top\right]\right\|_{\mathsf{op}} &\leq 2\sup_{v_k}\left\|\tilde{\phi}(Xv_k)E_k v_k^\top\right\|_{\mathsf{op}} \\
&= 2\sup_{v_k}\left(\left\|\tilde{\phi}(Xv_k)\right\|_2\|v_k\|_2 E_k\right) \\
&\leq 2C_1^2\sqrt{n},
\end{aligned} \tag{I.182}$$

because of (I.177). Then, by matrix Bernstein's inequality for rectangular matrices (see Exercise 5.4.15 in Vershynin [2018]), we have that

$$\mathbb{P}_V\left(\left\|\tilde{\Phi}EV - \mathbb{E}_V\left[\tilde{\Phi}EV\right]\right\|_{\mathsf{op}} \geq t\right) \leq (n+d)\exp\left(-\frac{t^2/2}{\sigma^2 + 2C_1^2\sqrt{n}t/3}\right), \tag{I.183}$$

where σ^2 is defined as

$$\sigma^2 = p\max\left(\left\|\mathbb{E}_{v_k}\left[W_k W_k^\top\right]\right\|_{\mathsf{op}}, \left\|\mathbb{E}_{v_k}\left[W_k^\top W_k\right]\right\|_{\mathsf{op}}\right). \tag{I.184}$$

For every matrix A, we have $\mathbb{E}\left[(A-\mathbb{E}[A])(A-\mathbb{E}[A])^\top\right] = \mathbb{E}\left[AA^\top\right] - \mathbb{E}[A]\mathbb{E}[A]^\top \preceq \mathbb{E}\left[AA^\top\right]$. Thus,

$$\begin{aligned}
\left\|\mathbb{E}_{v_k}\left[W_k W_k^\top\right]\right\|_{\mathsf{op}} &\leq \left\|\mathbb{E}_{v_k}\left[\tilde{\phi}(Xv_k)E_k v_k^\top v_k E_k \tilde{\phi}(Xv_k)^\top\right]\right\|_{\mathsf{op}} \\
&\leq \left\|\mathbb{E}_{v_k}\left[\tilde{\phi}(Xv_k)\tilde{\phi}(Xv_k)^\top\right]\right\|_{\mathsf{op}} \sup_{v_k}\left(E_k\left\|v_k\right\|_2^2\right) \\
&\leq C_1^2\left\|\mathbb{E}_{v_k}\left[\tilde{\phi}(Xv_k)\tilde{\phi}(Xv_k)^\top\right]\right\|_{\mathsf{op}} \\
&= \frac{C_1^2}{p}\left\|\mathbb{E}_V\left[\tilde{K}\right]\right\|_{\mathsf{op}} \\
&= O(1),
\end{aligned} \tag{I.185}$$

where the last step is a direct consequence of Lemma I.13, and holds with probability at least $1-2\exp\left(-c_4\log^2 n\right)$ over X. For the other argument in the $\max$ in (I.184), we similarly have

$$\begin{aligned}
\left\|\mathbb{E}\left[W_k^\top W_k\right]\right\|_{\mathsf{op}} &\leq \left\|\mathbb{E}_{v_k}\left[v_k E_k \tilde{\phi}(Xv_k)^\top \tilde{\phi}(Xv_k) E_k v_k^\top\right]\right\|_{\mathsf{op}} \\
&\leq \left\|\mathbb{E}_{v_k}\left[v_k v_k^\top\right]\right\|_{\mathsf{op}} \sup_{v_k}\left(E_k\left\|\tilde{\phi}(Xv_k)\right\|_2^2\right) \\
&\leq \frac{1}{d}C_1^2 n \\
&= O\left(\frac{n}{d}\right).
\end{aligned} \tag{I.186}$$

Then, plugging these last two equations in (I.183) we get

$$\begin{aligned}
&\mathbb{P}_V\left(\left\|\tilde{\Phi}EV - \mathbb{E}_V\left[\tilde{\Phi}EV\right]\right\|_{\mathsf{op}} \geq \sqrt{\frac{pn}{d}}\log n\right) \leq \\
&\quad\leq (n+d)\exp\left(-\frac{(pn/d)\log^2 n/2}{C_2(pn/d) + 2C_1^2\sqrt{np}\sqrt{n/d}\log n/3}\right) \\
&\quad\leq 2\exp\left(-c_5\log^2 n\right),
\end{aligned} \tag{I.187}$$

where we used Assumption 8.7. Then, applying the triangle inequality and using (I.180) and (I.187), we get

$$\left\|\tilde{\Phi}EV\right\|_{\mathsf{op}} = O\left(\sqrt{\frac{pn}{d}}\log n\right), \tag{I.188}$$

with probability at least $1-2\exp\left(-c_6\log^2 n\right)$ over X, V. To conclude, since $E = I$ with probability at least $1-2p\exp\left(-c_2 d\right)$, using Assumption 8.7 gives the desired result. □

Lemma I.20. *We have that*

$$\left\|V^\top P_\Lambda^\perp \Phi^+\right\|_{\mathsf{op}} = O\left(\frac{\sqrt{d}}{n} + \frac{\sqrt{n}\log^3 d}{d^{3/2}}\right), \tag{I.189}$$

with probability at least $1 - 2\exp\left(-c\log^2 n\right)$ *over* X *and* V*, where* c *is an absolute constant.*

Proof. Let $\tilde{\mu}^2 := \sum_{l=3}^{\infty} \mu_l^2$, where μ_l denotes the l-th Hermite coefficient of ϕ, and let $E \in \mathbb{R}^{n\times n}$ be the matrix defined as

$$E = K - p\left(\mu_1^2 \frac{XX^\top}{d} + \tilde{\mu}^2 I\right). \tag{I.190}$$

Note that

$$\begin{aligned}
\left\|EK^{-1}\right\|_{\mathsf{op}} &\le \left\|K - \mathbb{E}_V[K]\right\|_{\mathsf{op}} \left\|K^{-1}\right\|_{\mathsf{op}} + \left\|\mathbb{E}_V[K] - p\left(\mu_1^2\frac{XX^\top}{d} + \tilde{\mu}^2 I\right)\right\|_{\mathsf{op}} \left\|K^{-1}\right\|_{\mathsf{op}} \\
&= \left\|\mathbb{E}_V[K]^{-1/2}\left(K - \mathbb{E}_V[K]\right)\mathbb{E}_V[K]^{-1/2}\right\|_{\mathsf{op}} \left\|\mathbb{E}_V[K]\right\|_{\mathsf{op}} \left\|K^{-1}\right\|_{\mathsf{op}} \\
&\quad + O\left(p\frac{n\log^3 n}{d^{3/2}}\frac{1}{p}\right) \\
&= O\left(\sqrt{\frac{n}{p}}\log n \log p \frac{np}{d}\frac{1}{p}\right) + O\left(\frac{n\log^3 n}{d^{3/2}}\right) \\
&= O\left(\frac{\sqrt{n}}{d}\log^2 n + \frac{n\log^3 n}{d^{3/2}}\right) \\
&= O\left(\frac{n\log^3 n}{d^{3/2}}\right),
\end{aligned} \tag{I.191}$$

where in the second line we used (I.75) and $\lambda_{\min}(K) = \Omega(p)$, which holds by Lemma I.11; in the third line we used Lemma I.10, $\lambda_{\min}(K) = \Omega(p)$ and $\left\|\mathbb{E}_V[K]\right\|_{\mathsf{op}} = O(np/d)$, which follows from Weyl inequality applied to (I.70) and (I.76) (conditioning on $\|X\|_{\mathsf{op}} = O(\sqrt{n})$, given by Lemma I.8); in the fourth line we used $p = \Omega(n^2)$, and the last step holds due to Assumption 8.7. The full equation as a whole holds with probability at least $1 - 2\exp\left(-c_1\log^2 n\right)$ over X and V.

By the Woodbury matrix identity (or Hua's identity), we have

$$\begin{aligned}
K^{-1} &= \left(p\left(\mu_1^2\frac{XX^\top}{d} + \tilde{\mu}^2 I\right) + E\right)^{-1} \\
&= \left(\mu_1^2 p\frac{XX^\top}{d} + \tilde{\mu}^2 pI\right)^{-1} - \left(\mu_1^2 p\frac{XX^\top}{d} + \tilde{\mu}^2 pI\right)^{-1} EK^{-1}.
\end{aligned} \tag{I.192}$$

Then, since we have

$$\left\|\frac{pX^\top}{d}\left(\mu_1^2 p\frac{XX^\top}{d} + \tilde{\mu}^2 pI\right)^{-1}\right\|_{\mathsf{op}} \le \frac{1}{\mu_1^2\sqrt{\lambda_{\min}(X^\top X)}} = O\left(\frac{1}{\sqrt{n}}\right), \tag{I.193}$$

with probability at least $1 - 2\exp\left(-c_2 \log^2 n\right)$ over X due to Lemma I.8, (I.192) allows us to write

$$\left\| \frac{pX^\top}{d} \left(K^{-1} - \left(\mu_1^2 p \frac{XX^\top}{d} + \tilde{\mu}^2 pI \right)^{-1} \right) \right\|_{\mathsf{op}} = \left\| \frac{pX^\top}{d} \left(\mu_1^2 p \frac{XX^\top}{d} + \tilde{\mu}^2 pI \right)^{-1} EK^{-1} \right\|_{\mathsf{op}} = O\left(\frac{\sqrt{n} \log^3 n}{d^{3/2}} \right), \tag{I.194}$$

where we used (I.191) and (I.193) in the last step, which then holds with probability at least $1 - 2\exp\left(-c_3 \log^2 n\right)$ over X and V. Notice that, with this same probability, the application of the triangle inequality to (I.193) and (I.194), together with Assumption 8.7, gives

$$\left\| \frac{pX^\top}{d} K^{-1} \right\|_{\mathsf{op}} = O\left(\frac{1}{\sqrt{n}} \right). \tag{I.195}$$

Recalling that $\Phi = \mu_1 X V^\top + \tilde{\Phi}$, another triangle inequality yields

$$\begin{aligned} \left\| V^\top \Phi^\top K^{-1} - \frac{\mu_1 p X^\top}{d} K^{-1} \right\|_{\mathsf{op}} &\leq \left\| V^\top \tilde{\Phi}^\top K^{-1} \right\|_{\mathsf{op}} + \mu_1 \left\| \frac{d}{p} V^\top V - I \right\|_{\mathsf{op}} \left\| \frac{pX^\top}{d} K^{-1} \right\|_{\mathsf{op}} \\ &= O\left(\sqrt{\frac{pn}{d}} \log n \frac{1}{p} + \sqrt{\frac{d}{p}} \frac{1}{\sqrt{n}} \right) \\ &= O\left(\frac{\log n}{\sqrt{dn}} \right), \end{aligned} \tag{I.196}$$

where the second line holds with probability at least $1 - 2\exp\left(-c_4 \log^2 n\right)$ over X and V because of Lemma I.19, Theorem 4.6.1 in Vershynin [2018] and (I.195), and the last step is a consequence of Assumption 8.7.

We now define $\varrho_\Lambda \in \mathbb{R}^{n \times n}$ as the projector on the span of the eigenvectors associated with the d largest eigenvalues of K. This implies $P_\Lambda \Phi^\top = \Phi^\top \varrho_\Lambda$, and therefore $P_\Lambda^\perp \Phi^+ = \Phi^\top K^{-1} \varrho_\Lambda^\perp$. Hence, we have

$$\left\| \mu_1 V X^\top \varrho_\Lambda^\perp \right\|_{\mathsf{op}} \leq \left\| \Phi^\top \varrho_\Lambda^\perp \right\|_{\mathsf{op}} + \left\| \tilde{\Phi}^\top \varrho_\Lambda^\perp \right\|_{\mathsf{op}} \leq \sqrt{\lambda_{d+1}(K)} + \left\| \tilde{\Phi} \right\|_{\mathsf{op}} \left\| \varrho_\Lambda^\perp \right\|_{\mathsf{op}} = O\left(\sqrt{p} \right), \tag{I.197}$$

where the last step is a consequence of Lemma I.11 and Lemma I.14, and holds with probability at least $1 - 2\exp\left(-c_5 \log^2 n\right)$ over X and V. Then, conditioning on this high probability event and on the event $\lambda_{\min}\left(V^\top V\right) = \Omega(p/d)$, which holds with probability at least $1 - 2\exp\left(-c_6 d\right)$ over V due to Lemma I.8, we have

$$\left\| X^\top \varrho_\Lambda^\perp \right\|_{\mathsf{op}} \leq \frac{1}{\mu_1 \sqrt{\lambda_{\min}\left(V^\top V\right)}} \left\| \mu_1 V X^\top \varrho_\Lambda^\perp \right\|_{\mathsf{op}} = O\left(\sqrt{d} \right). \tag{I.198}$$

Thus, with probability at least $1-2\exp\left(-c_7\log^2 n\right)$ over X and V, we have

$$\begin{aligned}
\left\|V^\top P_\Lambda^\perp \Phi^+\right\|_{\mathsf{op}} &\le \left\|\frac{\mu_1 p X^\top}{d} K^{-1}\varrho_\Lambda^\perp\right\|_{\mathsf{op}} + \left\|V^\top\Phi^\top K^{-1} - \frac{\mu_1 p X^\top}{d}K^{-1}\right\|_{\mathsf{op}} \left\|\varrho_\Lambda^\perp\right\|_{\mathsf{op}} \\
&\le \left\|\frac{\mu_1 p X^\top}{d}\left(\mu_1^2 p\frac{XX^\top}{d} + \tilde{\mu}^2 pI\right)^{-1}\varrho_\Lambda^\perp\right\|_{\mathsf{op}} \\
&\quad + \left\|\frac{\mu_1 p X^\top}{d}\left(K^{-1} - \left(\mu_1^2 p\frac{XX^\top}{d} + \tilde{\mu}^2 pI\right)^{-1}\right)\right\|_{\mathsf{op}} \left\|\varrho_\Lambda^\perp\right\|_{\mathsf{op}} + O\left(\frac{\log n}{\sqrt{dn}}\right) \\
&= \left\|\left(\mu_1 X^\top X + \tilde{\mu}^2 dI/\mu_1\right)^{-1} X^\top \varrho_\Lambda^\perp\right\|_{\mathsf{op}} + O\left(\frac{\sqrt{n}\log^3 n}{d^{3/2}}\right) + O\left(\frac{\log n}{\sqrt{dn}}\right) \\
&\le \left\|\left(\mu_1 X^\top X + \tilde{\mu}^2 dI/\mu_1\right)^{-1}\right\|_{\mathsf{op}} \left\|X^\top \varrho_\Lambda^\perp\right\|_{\mathsf{op}} + O\left(\frac{\sqrt{n}\log^3 n}{d^{3/2}}\right) \\
&= O\left(\frac{1}{n}\sqrt{d}\right) + O\left(\frac{\sqrt{n}\log^3 n}{d^{3/2}}\right) \\
&= O\left(\frac{\sqrt{d}}{n} + \frac{\sqrt{n}\log^3 n}{d^{3/2}}\right),
\end{aligned} \tag{I.199}$$

where the second step holds due to (I.196), the third step due to (I.194), the fourth due to Assumption 8.7, and the fifth due to Lemma I.8 and (I.198). This, together with Assumption 8.7, provides the desired result. □

Lemma I.21. *Let Assumptions 8.5 and 8.6 hold, and let* $p=\omega(d)$ *and* $\log p=\Theta(\log n)=\Theta(\log d)$. *Let* $x\sim\mathcal{P}_X$. *Then, we have*

$$\left\|\mathbb{E}_x\left[\tilde{\varphi}(x)\tilde{\varphi}(x)^\top\right]\right\|_{\mathsf{op}} = O\left(\log^4 n + \frac{p\log^3 d}{d^{3/2}}\right), \tag{I.200}$$

with probability at least $1-2\exp\left(-c\log^2 n\right)$ *over* V, *where* c *is an absolute constant.*

Proof. Let N be a positive natural number that will be defined later, and let $\tilde{\Phi}_N\in\mathbb{R}^{N\times p}$ be a matrix containing $\tilde{\varphi}(\hat{x}_i)$ in its i-th row, where every $\{\hat{x}_i\}_{i=1}^N$ is sampled independently from $\mathcal{P}_X$. Importantly, $\hat{x}_i$ is different from x_i (as they are defined as auxiliary random variables only useful for the purposes of this proof), but $\tilde{\varphi}(\hat{x}_i)=\tilde{\phi}(V\hat{x}_i)$ is defined with the same random features V in (8.2). Set

$$n' = \min\left(\left\lfloor\frac{p}{\log^4 p}\right\rfloor, \left\lfloor\frac{d^{3/2}}{\log^3 d}\right\rfloor\right), \qquad N = p^2 n', \tag{I.201}$$

where n' is a positive integer (as p is large enough). Note that this definition guarantees that the triple (n',d,p) satisfies the scalings

$$n' = O\left(\frac{p}{\log^4 p}\right), \qquad n'\log^3 n' = O\left(d^{3/2}\right), \tag{I.202}$$

which provide the sufficient hypotheses to apply Lemma I.14 on a $n'\times p$ block of $\tilde{\Phi}_N$.

Then, $\tilde{\Phi}_N$ can be seen as the vertical stacking of p^2 matrices with size $n'\times p$. All these matrices are independent copies of each other (in the probability space of the $\hat{x}_i$-s), and each of these

has operator norm $O\left(\sqrt{p}\right)$ by Lemma I.14, with probability at least $1-2\exp\left(-c_1\log^2 n'\right)$. Thus, performing a union bound over these p^2 matrices, we get

$$\left\|\tilde{\Phi}_N^\top\tilde{\Phi}_N\right\|_{\mathsf{op}} = O\left(p^2\,p\right) = O\left(\frac{Np}{n'}\right) = O\left(N\log^4 p + \frac{Np\log^3 d}{d^{3/2}}\right), \tag{I.203}$$

with probability at least

$$1-2p^2\exp\left(-c_1\log^2 n'\right) \geq 1-2p^2\exp\left(-c_2\min\left(\log^2 p, \log^2 d\right)\right) \tag{I.204}$$

over V and $\{\hat{x}_i\}_{i=1}^N$. Note that the rows of $\tilde{\Phi}_N$ are identically distributed and such that

$$\sup_{\hat{x}_i}\|\tilde{\varphi}(\hat{x}_i)\|_2 \leq \left\|\tilde{\phi}(\mathbf{0})\right\|_2 + \tilde{L}\sup_{\hat{x}_i}\|V\hat{x}_i\|_2 \leq \left\|\tilde{\phi}(\mathbf{0})\right\|_2 + \tilde{L}\,\|V\|_{\mathsf{op}}\sup_{\hat{x}_i}\|\hat{x}_i\|_2 = O\left(\sqrt{p}\right), \tag{I.205}$$

where we denote with $\mathbf{0}\in\mathbb{R}^p$ a vector of zeros, $\tilde{L}$ is the Lipschitz constant of $\tilde{\phi}$, and the last step follows from the bound on $\|V\|_{\mathsf{op}}$ given by Lemma I.8, which holds with probability at least $1-2\exp\left(-c_3 d\right)$ over V (high probability event over which we will condition until the end of the proof). This readily gives, for some constant C_1,

$$\|\tilde{\varphi}(\hat{x}_i)\|_{\psi_2} \leq C_1\sup_{\hat{x}_i}\|\tilde{\varphi}(\hat{x}_i)\|_2 = O\left(\sqrt{p}\right). \tag{I.206}$$

We remark that the sub-Gaussian norm in (I.206) is intended in the probability space of $\hat{x}_i$, and that the bound holds jointly for every $i\in[N]$. Then, there exists a sufficiently small absolute constant C_2 such that $C_2\tilde{\Phi}_N/\sqrt{p}$ is a matrix with independent sub-Gaussian rows, with unit sub-Gaussian norm. Then, by Theorem 5.39 in Vershynin [2012] (see their Remark 5.40 and Equation (5.25)), we have that

$$\frac{C_2^2}{p}\left\|\frac{\tilde{\Phi}_N^\top\tilde{\Phi}_N}{N} - \mathbb{E}_x\left[\tilde{\varphi}(x)\tilde{\varphi}(x)^\top\right]\right\|_{\mathsf{op}} = O\left(\sqrt{\frac{p}{N}}\right), \tag{I.207}$$

with probability at least $1-2\exp\left(-c_4 p\right)$ over $\{\hat{x}_i\}_{i=1}^N$.

Then, we have

$$\begin{aligned}
\left\|\mathbb{E}_x\left[\tilde{\varphi}(x)\tilde{\varphi}(x)^\top\right]\right\|_{\mathsf{op}} &\leq \left\|\frac{\tilde{\Phi}_N^\top\tilde{\Phi}_N}{N} - \mathbb{E}_x\left[\tilde{\varphi}(x)\tilde{\varphi}(x)^\top\right]\right\|_{\mathsf{op}} + \frac{\left\|\tilde{\Phi}_N^\top\tilde{\Phi}_N\right\|_{\mathsf{op}}}{N} \\
&= O\left(p\sqrt{\frac{p}{N}}\right) + O\left(\log^4 p + \frac{p\log^3 d}{d^{3/2}}\right) \\
&= O\left(\sqrt{p}\sqrt{\frac{\log^4 p}{p}} + \sqrt{p}\sqrt{\frac{\log^3 d}{d^{3/2}}}\right) + O\left(\log^4 p + \frac{p\log^3 d}{d^{3/2}}\right) \\
&= O\left(\log^4 p + \frac{p\log^3 d}{d^{3/2}}\right),
\end{aligned} \tag{I.208}$$

where the first step follows from the triangle inequality, the second step is a consequence of (I.207) and (I.203), and the third step is a consequence of (I.201). Taking the intersection between the high probability events in (I.203) and (I.207), we have

$$\left\|\mathbb{E}_x\left[\tilde{\varphi}(x)\tilde{\varphi}(x)^\top\right]\right\|_{\mathsf{op}} = O\left(\log^4 n + \frac{p\log^3 d}{d^{3/2}}\right), \tag{I.209}$$

with probability at least $1 - 2p^2 \exp\left(-c_5 \min\left(\log^2 p, \log^2 d\right)\right) \geq 1 - 2\exp\left(-c_6 \log^2 n\right)$ over $\{\hat{x}_i\}_{i=1}^N$ and V. Note that the LHS of the previous equation does not depend on $\{\hat{x}_i\}_{i=1}^N$, which were introduced as auxiliary random variables. Thus, the high probability bound holds restricted to the probability space of V, and the desired result follows. □

Lemma I.22. *Let Assumptions 8.5 and 8.6 hold, and let* $n = o\left(p/\log^4 p\right)$, $n \log^3 n = o\left(d^{3/2}\right)$, $n = \omega(d)$ *and* $\log n = \Theta(\log p)$. *Then, we have*

$$\mathbb{E}_x\left[\left(\varphi(x)^\top \theta^*\right)^2\right] = O\left(1\right), \tag{I.210}$$

with probability at least $1 - 2\exp\left(-c \log^2 n\right)$ *over* X *and* V, *where* c *is an absolute constant.*

Proof. First, we can upper bound the LHS of (I.210) as

$$\mathbb{E}_x\left[\left(\varphi(x)^\top \theta^*\right)^2\right] \leq 2\mu_1^2 \mathbb{E}_x\left[\left(x^\top V^\top \theta^*\right)^2\right] + 2\mathbb{E}_x\left[\left(\tilde{\varphi}(x)^\top \theta^*\right)^2\right]. \tag{I.211}$$

We bound the two terms separately. Since x is distributed according to $\mathcal{P}_X$, it is sub-Gaussian with $\|x\|_{\psi_2} = O\left(1\right)$. Then, we can bound its second moment (see Vershynin [2018], Proposition 2.5.2) as

$$\mathbb{E}_x\left[\left(x^\top V^\top \theta^*\right)^2\right] \leq C\left\|V^\top \theta^*\right\|_2^2 \leq C\left\|V^\top \Phi^+\right\|_{\mathsf{op}}^2 \|Y\|_2^2 = O\left(\frac{1}{n} n\right) = O\left(1\right), \tag{I.212}$$

where the third step holds with probability at least $1 - 2\exp\left(-c_1 \log^2 n\right)$ over X and V because of Lemma I.16. Next, we bound the second term of (I.211) as

$$\begin{aligned}
\mathbb{E}_x\left[\left(\tilde{\varphi}(x)^\top \theta^*\right)^2\right] &= \left(\theta^*\right)^\top \mathbb{E}_x\left[\tilde{\varphi}(x)\tilde{\varphi}(x)^\top\right] \theta^* \\
&\leq \left\|\mathbb{E}_x\left[\tilde{\varphi}(x)\tilde{\varphi}(x)^\top\right]\right\|_{\mathsf{op}} \|\theta^*\|_2^2 \\
&\leq \left\|\mathbb{E}_x\left[\tilde{\varphi}(x)\tilde{\varphi}(x)^\top\right]\right\|_{\mathsf{op}} \left\|\Phi^+\right\|_{\mathsf{op}}^2 \|Y\|_2^2 \\
&= O\left(\left(\log^4 n + \frac{p \log^3 d}{d^{3/2}}\right) \frac{1}{p} n\right) \\
&= O\left(\frac{n \log^4 n}{p} + \frac{n \log^3 d}{d^{3/2}}\right) = o\left(1\right),
\end{aligned} \tag{I.213}$$

where the fourth line holds because of Lemmas I.21 and I.11 with probability at least $1 - 2\exp\left(-c_2 \log^2 n\right)$ over X and V. The last line provides the desired result. □

APPENDIX J

Appendix for Chapter 9

J.1 Proofs on differential privacy

The notion of zCDP can be converted to (ε, δ)-DP, which in turn is defined below.

Definition J.1 ((ε, δ)-DP [Dwork et al., 2006]). *A randomized algorithm $\mathcal{A}$ satisfies (ε, δ)-differential privacy if for any pair of adjacent datasets D, D', and for any subset of the parameters space $S \subseteq \mathbb{R}^d$, we have*

$$\mathbb{P}\left(\mathcal{A}(D) \in S\right) \leq e^{\varepsilon}\mathbb{P}\left(\mathcal{A}(D') \in S\right) + \delta. \tag{J.1}$$

For completeness, we also provide the following definition.

Definition J.2 (Rényi DP [Mironov, 2017]). *Given $\alpha \in (1, +\infty)$ and $\varepsilon \geq 0$, an algorithm $\mathcal{A}$ satisfies (α, ε) Rényi DP if for any pair of adjacent datasets D, D' we have $D_\alpha\left(\mathcal{A}(D) \,\|\, \mathcal{A}(D')\right) \leq \varepsilon$, where $D_\alpha\left(\mathcal{A}(D) \,\|\, \mathcal{A}(D')\right)$ is the Rényi Divergence [Rényi, 1961] between the probability distributions induced by the randomness of $\mathcal{A}$, i.e.,*

$$D_\alpha\left(\mathcal{A}(D) \,\|\, \mathcal{A}(D')\right) = \frac{1}{\alpha - 1} \ln \int \left(\frac{p_{\mathcal{A}(D)}(\theta)}{p_{\mathcal{A}(D')}(\theta)}\right)^{\alpha} p_{\mathcal{A}(D')}(\theta) \mathrm{d}\theta. \tag{J.2}$$

Note that Definitions 9.1 and J.2 imply that an algorithm is ρ-zCDP if, for any $\alpha > 1$, it is also $(\alpha, \rho\alpha)$ Rényi DP. The following proposition allows to translate Rényi DP and zCDP to (ε, δ)-DP.

Proposition J.3 (Proposition 1.3 in Bun and Steinke [2016]). *If $\mathcal{A}$ satisfies $\rho^2/2$-zCDP, it also satisfies $\left(\rho^2/2 + \rho\sqrt{2\ln(1/\delta)}, \delta\right)$-DP, for any $\delta \in (0, 1)$.*

Then, if we consider δ such that $\rho \leq \sqrt{\ln(1/\delta)}$, we achieve (ε, δ)-DP if we have

$$\rho^2/2 + \rho\sqrt{2\ln(1/\delta)} \leq 2\rho\sqrt{\ln(1/\delta)} \leq \varepsilon, \tag{J.3}$$

which means that for algorithms respecting $\rho^2/2$-zCDP, we can replace 2ρ by $\varepsilon/\sqrt{\ln(1/\delta)}$ in the error bounds to evaluate the cost of privacy in terms of (ε, δ)-DP.

J.1.1 Proof of Proposition 9.2

Let us define the family of functions $\ell_{k,C_{\mathsf{clip}}}(\cdot) : \mathbb{R} \to \mathbb{R}$, for all $k \in [n]$, such that $\ell_{k,C_{\mathsf{clip}}}(0) = 0$, and

$$\ell'_{k,C_{\mathsf{clip}}}(z) = z \min\left(1, \frac{C_{\mathsf{clip}}}{|z| \left\|x_k\right\|_2}\right). \tag{J.4}$$

In words, $\ell_{k,C_{\mathsf{clip}}}(z)$ is a quadratic function, but linearized for sufficiently large values of $|z|$, such that it is $C_{\mathsf{clip}}/\left\|x_k\right\|_2$-Lipschitz. Then, we have

$$\begin{aligned}
\bar{g}_k &= g_k \min\left(1, \frac{C_{\mathsf{clip}}}{\left\|g_k\right\|_2}\right) \\
&= x_k \left(x_k^\top \theta_{k-1} - y_k\right) \min\left(1, \frac{C_{\mathsf{clip}}}{\left\|x_k\right\|_2 \left|x_k^\top \theta_{k-1} - y_k\right|}\right) \\
&= x_k \ell'_{k,C_{\mathsf{clip}}}\left(x_k^\top \theta_{k-1} - y_k\right) \\
&= \nabla_\theta \ell_{k,C_{\mathsf{clip}}}\left(x_k^\top \theta_{k-1} - y_k\right),
\end{aligned} \tag{J.5}$$

where the first step follows from the definition of $\bar{g}_k$ in Algorithm 9.1. Furthermore, for any $\theta, \theta' \in \mathbb{R}^d$, we have

$$\begin{aligned}
&\left\|\nabla_\theta \ell_{k,C_{\mathsf{clip}}}(x_k^\top \theta - y_k) - \nabla_\theta \ell_{k,C_{\mathsf{clip}}}(x_k^\top \theta' - y_k)\right\|_2 \\
&\quad = \left\|x_k\right\|_2 \left|\ell'_{k,C_{\mathsf{clip}}}(x_k^\top \theta - y_k) - \ell'_{k,C_{\mathsf{clip}}}(x_k^\top \theta' - y_k)\right| \\
&\quad \leq \left\|x_k\right\|_2 \left|x_k^\top (\theta - \theta')\right| \\
&\quad \leq \left\|x_k\right\|_2^2 \left\|\theta - \theta'\right\|_2,
\end{aligned} \tag{J.6}$$

where the second step follows from the fact that $\ell'_{k,C_{\mathsf{clip}}}(z)$ is a 1-Lipschitz function. Let us now define

$$\bar{\ell}_{k,C_{\mathsf{clip}}}(z) = \min\left(1, \frac{2}{\left\|x_k\right\|_2^2 \eta_k}\right) \ell_{k,C_{\mathsf{clip}}}(z). \tag{J.7}$$

Then, we have that every iteration of Algorithm 9.1 takes the form

$$\theta_k = \theta_{k-1} - \eta_k \nabla_\theta \bar{\ell}_{k,C_{\mathsf{clip}}}(x_k^\top \theta_{k-1} - y_k) + 2C_{\mathsf{clip}} \sigma_k b_k, \tag{J.8}$$

where $\bar{\ell}_{k,C_{\mathsf{clip}}}(x_k^\top \theta - y_k)$ is C_{clip}-Lipschitz with respect to θ, and it is $2/\eta_k$-smooth, i.e.,

$$\left\|\nabla_\theta \bar{\ell}_{k,C_{\mathsf{clip}}}(x_k^\top \theta - y_k) - \nabla_\theta \bar{\ell}_{k,C_{\mathsf{clip}}}(x_k^\top \theta' - y_k)\right\|_2 \leq \frac{2}{\eta_k} \left\|\theta - \theta'\right\|_2, \tag{J.9}$$

due to (J.6) and (J.7). Thus, the desired result follows from Theorem 3.1 in Feldman et al. [2020], after setting their batch sizes $\{B_k\}$ identically equal to 1, and their projection set $\mathcal{K}$ equal to all $\mathbb{R}^d$. □

J.2 Auxiliary lemmas on $\mu_c(\theta)$ and $\nu_c(\theta)$

Lemma J.4. *Let Assumption 9.3 hold, and let $\mu_c(\theta)$ and $\nu_c(\theta)$ be defined according to* (9.14)*. Then, we have*

$$\mu_c(\theta) = \operatorname{erf}\left(\frac{c}{2\sqrt{\mathcal{P}(\theta)}}\right), \tag{J.10}$$

$$\nu_c(\theta) = \frac{c^2}{2\mathcal{P}(\theta)}\left(1 - \operatorname{erf}\left(\frac{c}{2\sqrt{\mathcal{P}(\theta)}}\right)\right) + F\left(\frac{c}{\sqrt{2\mathcal{P}(\theta)}}\right), \tag{J.11}$$

where

$$\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2}\mathsf{d}t, \qquad F(z) = \frac{1}{\sqrt{2\pi}}\int_{-z}^{z} t^2 e^{-t^2/2}\mathsf{d}t = \operatorname{erf}\left(\frac{z}{\sqrt{2}}\right) - \sqrt{\frac{2}{\pi}} z e^{-z^2/2}. \tag{J.12}$$

In particular, $\mu_c(\theta)$ and $\nu_c(\theta)$ depend only on c and the test risk $\mathcal{P}(\theta)$ via the ratio $c/\sqrt{2\mathcal{P}(\theta)}$.

Proof. Recall that

$$r(\theta, x, y) = x^\top\theta - y, \qquad r_c(\theta, x, y) = r(\theta, x, y)\min\left(1, \frac{c}{|r(\theta, x, y)|}\right), \tag{J.13}$$

$$\mu_c(\theta) = \frac{\left\|\mathbb{E}_{x,y}\left[r_c(\theta, x, y)\, x\right]\right\|_2}{\left\|\mathbb{E}_{x,y}\left[r(\theta, x, y)\, x\right]\right\|_2}, \qquad \nu_c(\theta) = \frac{\mathbb{E}_{x,y}\left[r_c(\theta, x, y)^2\right]}{\mathbb{E}_{x,y}\left[r(\theta, x, y)^2\right]}. \tag{J.14}$$

Until the end of the proof, we will use the notation $\mathsf{clip}_c(\cdot) : \mathbb{R} \to \mathbb{R}$ to denote the function such that

$$\mathsf{clip}_c(a) = a\min\left(1, \frac{c}{|a|}\right). \tag{J.15}$$

In particular, $r_c(\theta, x, y) = \mathsf{clip}_c\left(r(\theta, x, y)\right)$.

Let us look at the first entry of the vector $\mathbb{E}_{x,y}\left[r_c(\theta, x, y)\, x\right]$,

$$\mathbb{E}_{x,y}\left[r_c(\theta, x, y) x^\top e_1\right] = \mathbb{E}_{\rho_1,\rho_2}\left[\mathsf{clip}_c(\rho_1)\rho_2\right], \tag{J.16}$$

where the second step introduced ρ_1 and ρ_2, defined as two mean-0 Gaussian random variables, such that

$$\operatorname{Var}(\rho_1) = \left\|\Sigma^{1/2}(\theta - \theta^*)\right\|_2^2 + \zeta^2, \qquad \operatorname{Var}(\rho_2) = \Sigma_{11}, \qquad \operatorname{Cov}(\rho_1, \rho_2) = e_1^\top\Sigma(\theta - \theta^*). \tag{J.17}$$

Then, we have

$$\begin{aligned}
\mathbb{E}_{\rho_1,\rho_2}\left[\mathsf{clip}_c(\rho_1)\rho_2\right] &= \frac{\operatorname{Cov}(\rho_1, \rho_2)}{\operatorname{Var}(\rho_1)}\mathbb{E}_{\rho_1}\left[\mathsf{clip}_c(\rho_1)\rho_1\right] \\
&= \frac{\operatorname{Cov}(\rho_1, \rho_2)}{\sqrt{\operatorname{Var}(\rho_1)}}\mathbb{E}_{\hat{\rho}}\left[\mathsf{clip}_c(\sqrt{\operatorname{Var}(\rho_1)}\hat{\rho})\hat{\rho}\right] \\
&= \frac{e_1^\top\Sigma(\theta - \theta^*)}{\sqrt{2\mathcal{P}(\theta)}}\mathbb{E}_{\hat{\rho}}\left[\mathsf{clip}_c(\sqrt{2\mathcal{P}(\theta)}\hat{\rho})\hat{\rho}\right],
\end{aligned} \tag{J.18}$$

where we used $\mathcal{P}(\theta) = \mathcal{R}(\theta) + \zeta^2/2 = \mathrm{Var}(\rho_1)/2$ and we introduced the standard Gaussian random variable $\hat{\rho}$. As this argument holds for any component of the vector $\mathbb{E}_{x,y}\left[r_c(\theta, x, y)\, x\right]$, plugging the equation above in (J.16) gives

$$\begin{aligned}\left\|\mathbb{E}_{x,y}\left[r_c(\theta, x, y)\, x\right]\right\|_2 &= \frac{\left\|\Sigma(\theta - \theta^*)\right\|_2}{\sqrt{2\mathcal{P}(\theta)}}\mathbb{E}_{\hat{\rho}}\left[\mathsf{clip}_c(\sqrt{2\mathcal{P}(\theta)}\hat{\rho})\hat{\rho}\right] \\ &= \frac{\left\|\mathbb{E}_{x,y}\left[r(\theta, x, y)\, x\right]\right\|_2}{\sqrt{2\mathcal{P}(\theta)}}\mathbb{E}_{\hat{\rho}}\left[\mathsf{clip}_c(\sqrt{2\mathcal{P}(\theta)}\hat{\rho})\hat{\rho}\right],\end{aligned} \tag{J.19}$$

where in the second step we used that $\mathbb{E}_{x,y}\left[r(\theta, x, y)\, x\right] = \Sigma(\theta - \theta^*)$. Then, we also have

$$\mu_c(\theta) = \frac{\mathbb{E}_{\hat{\rho}}\left[\mathsf{clip}_c(\sqrt{2\mathcal{P}(\theta)}\hat{\rho})\hat{\rho}\right]}{\sqrt{2\mathcal{P}(\theta)}}. \tag{J.20}$$

Defining the shorthand $c'(\theta) = c/\sqrt{2\mathcal{P}(\theta)}$, the numerator of the expression above yields

$$\begin{aligned}\mathbb{E}_{\hat{\rho}}\left[\mathsf{clip}_c(\sqrt{2\mathcal{P}(\theta)}\hat{\rho})\hat{\rho}\right] &= \frac{\sqrt{2\mathcal{P}(\theta)}}{\sqrt{2\pi}}\int_{-c'(\theta)}^{c'(\theta)}\hat{\rho}^2 e^{-\hat{\rho}^2/2}\mathsf{d}\hat{\rho} + \frac{2c}{\sqrt{2\pi}}\int_{c'(\theta)}^{+\infty}\hat{\rho}e^{-\hat{\rho}^2/2}\mathsf{d}\hat{\rho} \\ &= \frac{\sqrt{2\mathcal{P}(\theta)}}{\sqrt{2\pi}}\left(-\hat{\rho}e^{-\hat{\rho}^2/2}\Big|_{-c'(\theta)}^{c'(\theta)} + \int_{-c'(\theta)}^{c'(\theta)} e^{-\hat{\rho}^2/2}\mathsf{d}\hat{\rho}\right) - \frac{2c}{\sqrt{2\pi}}\, e^{-\hat{\rho}^2/2}\Big|_{c'(\theta)}^{+\infty} \\ &= \frac{\sqrt{2\mathcal{P}(\theta)}}{\sqrt{2\pi}}\left(-2c'(\theta)e^{-c'(\theta)^2/2} + \int_{-c'(\theta)}^{c'(\theta)} e^{-\hat{\rho}^2/2}\mathsf{d}\hat{\rho}\right) + \frac{2c}{\sqrt{2\pi}}e^{-c'(\theta)^2/2} \\ &= \frac{\sqrt{\mathcal{P}(\theta)}}{\sqrt{\pi}}\int_{-c'(\theta)}^{c'(\theta)} e^{-\hat{\rho}^2/2}\mathsf{d}\hat{\rho} \\ &= \frac{2\sqrt{2}\sqrt{\mathcal{P}(\theta)}}{\sqrt{\pi}}\int_0^{c'(\theta)/\sqrt{2}} e^{-\hat{\rho}^2}\mathsf{d}\hat{\rho} \\ &= \sqrt{2\mathcal{P}(\theta)}\,\mathrm{erf}\left(\frac{c}{\sqrt{4\mathcal{P}(\theta)}}\right),\end{aligned} \tag{J.21}$$

which, plugged in (J.20), gives the first part of the thesis.

For the second part of the thesis, following the same argument we used to write (J.20), we have

$$\nu_c(\theta) = \frac{\mathbb{E}_{\hat{\rho}}\left[\mathsf{clip}_c(\sqrt{2\mathcal{P}(\theta)}\hat{\rho})^2\right]}{2\mathcal{P}(\theta)}, \tag{J.22}$$

where, as before, $\hat{\rho}$ denotes a standard Gaussian random variable. Then, we have

$$\begin{aligned}\nu_c(\theta) &= \frac{1}{\sqrt{2\pi}}\int_{-c'(\theta)}^{c'(\theta)}\hat{\rho}^2 e^{-\hat{\rho}^2/2}\mathsf{d}\hat{\rho} + \frac{2c'(\theta)^2}{\sqrt{2\pi}}\int_{c'(\theta)}^{+\infty} e^{-\hat{\rho}^2/2}\mathsf{d}\hat{\rho} \\ &= \frac{1}{\sqrt{2\pi}}\int_{-c'(\theta)}^{c'(\theta)}\hat{\rho}^2 e^{-\hat{\rho}^2/2}\mathsf{d}\hat{\rho} + c'(\theta)^2\left(1 - \frac{2}{\sqrt{2\pi}}\int_0^{c'(\theta)} e^{-\hat{\rho}^2/2}\mathsf{d}\hat{\rho}\right) \\ &= \frac{c^2}{2\mathcal{P}(\theta)}\left(1 - \mathrm{erf}\left(\frac{c}{2\sqrt{\mathcal{P}(\theta)}}\right)\right) + F(c'(\theta)),\end{aligned} \tag{J.23}$$

which concludes the proof. □

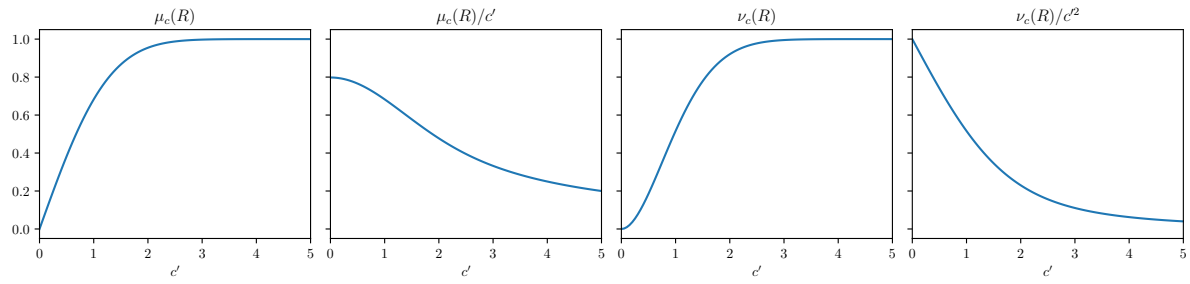


Figure J.1: The functions $\mu_c(R)$, $\mu_c(R)/c'$, $\nu_c(R)$, $\nu_c(R)/c'$, plotted as a function of $c' = c/\sqrt{2R+\zeta^2}$.

Lemma J.5. *Let Assumption 9.3 hold, and let $\mu_c(R)$ and $\nu_c(R)$ be defined according to (9.14), where R denotes a generic value of the test risk, since $\mu_c(\theta)$ and $\nu_c(\theta)$ depend only on c and the test risk $\mathcal{P}(\theta)$ due to Lemma J.4. Then, for any $c > 0$, we have that*

$$\underline{c}_\mu(c,\zeta) < \frac{\mu_c(R)\sqrt{2R+\zeta^2}}{c} < \sqrt{\frac{2}{\pi}}, \qquad \underline{c}_\nu(c,\zeta) < \frac{\nu_c(R)(2R+\zeta^2)}{c^2} < 1, \tag{J.24}$$

where $\underline{c}_\mu(c,\zeta)$ and $\underline{c}_\nu(c,\zeta)$ denote two positive constants which depend on the values of c and ζ and are monotonously decreasing in c.

We also have that, as $c/\zeta \to 0$,

$$\left|\frac{\mu_c(R)\sqrt{2R+\zeta^2}}{c} - \sqrt{\frac{2}{\pi}}\right| = o_{c/\zeta}(1), \qquad \left|\frac{\nu_c(R)\,(2R+\zeta^2)}{c^2} - 1\right| = o_{c/\zeta}(1). \tag{J.25}$$

Furthermore, we have

$$\begin{aligned} &\frac{\nu_c(R)(2R+\zeta^2)}{c^2} > \frac{1}{2}, && \text{if } \frac{c}{\sqrt{2R+\zeta^2}} \le 1, \\ &\nu_c(R) > \frac{1}{2}, && \text{if } \frac{c}{\sqrt{2R+\zeta^2}} > 1. \end{aligned} \tag{J.26}$$

Proof. Note that, introducing the notation

$$c' = \frac{c}{\sqrt{2R+\zeta^2}} \le \frac{c}{\zeta}, \tag{J.27}$$

we have that

$$\mu_c(R) = \operatorname{erf}\left(c'/\sqrt{2}\right) < 1, \tag{J.28}$$

and

$$\nu_c(R) = (c')^2\left(1 - \operatorname{erf}\left(c'/\sqrt{2}\right)\right) + F(c') < 1, \tag{J.29}$$

where the last inequalities can be verified directly via the definitions in (9.14). Furthermore, we have that both $\mu_c(R)$ and $\nu_c(R)$ are increasing functions of c', equal to 0 when $c' = 0$. This follows from the definition for $\mu_c(R)$, and can be promptly verified for $\nu_c(R)$ via derivation.

We further have that, for $c' > 0$,

$$\frac{\mu_c(R)\sqrt{2R+\zeta^2}}{c} = \frac{1}{c'}\operatorname{erf}\left(c'/\sqrt{2}\right) < \sqrt{\frac{2}{\pi}}, \tag{J.30}$$

and

$$\frac{\nu_c(R)\,(2R+\zeta^2)}{c^2} = \left(1 - \operatorname{erf}\left(c'/\sqrt{2}\right)\right) + \frac{F(c')}{(c')^2} < 1, \tag{J.31}$$

where both the LHSs are decreasing functions of c', going to 0 for $c' \to +\infty$ (this can be seen via explicit derivation with respect to c', and via the identity $0 \geq \int_0^z -2t^2 e^{-t^2}\mathrm{d}t = ze^{-z^2} - \int_0^z e^{-t^2}\mathrm{d}t$), and where the last inequalities can be verified computing the limit for $c' \to 0^+$ via l'Hôpital rule, which gives

$$\lim_{c'\to 0^+} \frac{\mu_c(R)\sqrt{2R+\zeta^2}}{c} = \sqrt{\frac{2}{\pi}}, \qquad \lim_{c'\to 0^+} \frac{\nu_c(R)\,(2R+\zeta^2)}{c^2} = 1. \tag{J.32}$$

Note that, as $R \geq 0$, the above limit is achieved when $c/\zeta \to 0$. Then, due to the inequality in (J.27), the first and second part of the thesis follow.

Note that, for $c' = 1$, we have

$$\nu_c(R) = 1 - \sqrt{\frac{2}{\pi e}} \approx 0.516. \tag{J.33}$$

Thus, the third and fourth part of the thesis follow from the monotonicity of $\nu_c(R)/c'$ and $\nu_c(R)$. □

J.3 Proofs for Section 9.4

We will use the notation $\|Z\|_{\psi_p} = \inf\{t > 0 : \mathbb{E}\exp(|Z|^p/t^p) \leq 2\}$ to denote the Orlicz norm of a random variable Z for any $p \geq 1$. We denote the inner product between two vectors a and b as $\langle a, b\rangle = a^\top b$ (with notation applicable to matrices $\langle aa^\top, A\rangle = a^\top A a$). Given two real-valued quantities a, b, we denote by $a \wedge b = \min(a, b)$. Given a quadratic functions $q : \mathbb{R}^d \to \mathbb{C}$ we define its $\|\cdot\|_{C^2}$ norm such that $\|q\|_{C^2} = \|\nabla^2 q\|_{\mathsf{op}} + \|\nabla q(0)\|_2 + |q(0)|$. Furthermore, we will denote with γ_n the ratio d/n for a fixed value of n.

We start by proving a general result about homogenized DP-GD, informally defined in Remark 9.7 and formally defined below.

Definition J.6 (Homogenized DP-GD)**.** *For any $t \in [0, 1)$, we define* homogenized DP-GD *(H-DP-GD) as the solution of the SDE*

$$\mathrm{d}\Theta_t = -2\bar{\eta}(t)\mu_c(\Theta_t)\nabla\mathcal{P}(\Theta_t)\mathrm{d}t + \bar{\eta}(t)\sqrt{\frac{2\nu_c(\Theta_t)\mathcal{P}(\Theta_t)\Sigma}{n}}\mathrm{d}B_t^s + 2\frac{\sqrt{d}}{n}c\tilde{\sigma}(t)\mathrm{d}B_t^p, \tag{J.34}$$

where $\Theta_0 = \theta_0 = 0$, B_t^s and B_t^p are two independent standard Brownian motions in $\mathbb{R}^d$, $\bar{\eta}(t) = \min(\tilde{\eta}(t), 2n/d)$, and $\tilde{\sigma}(t)$ is such that $\rho^2\tilde{\sigma}^2(t) = -\mathrm{d}\tilde{\eta}^2(t)/\mathrm{d}t$.

Theorem J.7 below shows that DP-GD is close to H-DP-GD over the class of quadratic functions Q, given by

$$Q = \{q : \mathbb{R}^d \to \mathbb{C} \;\text{ s.t . } q(v) = v^\top R(z;\Sigma)v/2, \;\text{ with }\; z \in \Omega\}, \tag{J.35}$$

where $R(z;\Sigma) = (\Sigma - zI)^{-1}$ is the resolvent matrix of Σ and $\Omega := \{w \in \mathbb{C} : |w| = 2\|\Sigma\|_{\mathsf{op}}\}$.

Theorem J.7. *Let Assumptions 9.3 and 9.4 hold. Let $\rho = \Theta(1)$, $n, d \to \infty$ s.t. $d/n \to \gamma \in (0, \infty)$. Denote by Θ_t and θ_k independent realizations of H-DP-GD (as per Definition J.6) and Algorithm 9.1. Then, with overwhelming probability, we have*

$$\sup_{t\in[0,1)} \left| q(\Theta_t - \theta^*) - q(\theta_{\lfloor tn \rfloor} - \theta^*) \right| = O\left(\frac{\log^2 n}{\sqrt{n}}\right), \tag{J.36}$$

uniformly for all functions $q \in Q$.

Proof. Consider the notation $u_k = \theta_k - \theta^*$. We have the following update rule for $k \in [n]$:

$$u_k = u_{k-1} - \bar{\eta}_k \bar{g}_k + 2c\sqrt{d}\sigma_k b_k, \tag{J.37}$$

where we recall that

$$\bar{\eta}_k = \min\left(\eta_k, \frac{2}{\|x_k\|_2}\right), \qquad \bar{g}_k = g_k \min\left(1, \frac{c\sqrt{d}}{\|g_k\|_2}\right), \qquad g_k = (\langle x_k, u_k\rangle - z_k)\, x_k. \tag{J.38}$$

Let $q : \mathbb{R}^d \to \mathbb{R}$ be a generic quadratic function. Then, using a Taylor expansion, the update rule for $q(u_k)$ reads

$$\begin{aligned} q(u_{k+1}) = q(u_k) &- \bar{\eta}_k \bar{g}_k^\top \nabla q(u_k) + 2c\sqrt{d}\sigma_k b_k^\top \nabla q(u_k) \\ &+ \frac{1}{2}\operatorname{tr}\left((2c\sqrt{d}\sigma_k b_k - \bar{\eta}_k \bar{g}_k)^{\otimes 2} \nabla^2 q(u_k)\right). \end{aligned} \tag{J.39}$$

Defining the σ-algebra $\mathcal{F}_k := \sigma(\{u_i\}_{i=0}^k)$ generated by the iterates of DP-GD in (J.39), we write the Doob's decomposition of the above process:

$$\begin{aligned} q(u_{k+1}) - q(u_k) = &-\frac{\bar{\eta}(k/n)}{n}\mu_c(u_k)\langle \Sigma u_k, \nabla q(u_k)\rangle \\ &+ \frac{\bar{\eta}^2(k/n)}{2n}\nu_c(u_k)\mathcal{P}(u_k)\frac{1}{n}\operatorname{tr}(\Sigma\nabla^2 q(u_k)) + \frac{1}{n}\frac{2d}{n^2}c^2\tilde{\sigma}(k/n)^2\operatorname{tr}(\nabla^2 q(u_k)) \\ &+ \Delta\mathcal{M}_k^{\mathsf{Grad}}(q) + \Delta\mathcal{M}_k^{\mathsf{Hess}}(q) + \mathbb{E}[\Delta\mathcal{E}_k^{\mathsf{Hess}}(q) \mid \mathcal{F}_k] \\ &+ \mathcal{M}_k^{\mathsf{Noise}}(q) + \mathbb{E}[\Delta\mathcal{E}_k^{\mathsf{Noise}}(q) \mid \mathcal{F}_k] + \mathbb{E}[\Delta\mathcal{E}_k^{\mathsf{Step}}(q) \mid \mathcal{F}_k], \end{aligned} \tag{J.40}$$

where we recall the notation $\bar{\eta}(k/n) = \min(\tilde{\eta}(k/n), 2/\gamma_n)$, $\eta_k = \tilde{\eta}(k/n)/n$, and where we

introduced the shorthands

$$\begin{aligned}
\Delta\mathcal{M}_k^{\mathsf{Grad}}(q) &:= -\bar{\eta}_k\langle\bar{g}_k, \nabla q(u_k)\rangle + \bar{\eta}_k\mu_c(u_k)\langle\Sigma u_k, \nabla q(u_k)\rangle \\
\Delta\mathcal{M}_k^{\mathsf{Hess}}(q) &:= \frac{1}{2}\langle(\bar{\eta}_k\bar{g}_k)^{\otimes 2}, \nabla^2 q(u_k)\rangle - \frac{1}{2}\langle\mathbb{E}[(\bar{\eta}_k\bar{g}_k)^{\otimes 2} \mid \mathcal{F}_k], \nabla^2 q(u_k)\rangle \\
\mathbb{E}[\Delta\mathcal{E}_k^{\mathsf{Hess}}(q) \mid \mathcal{F}_k] &:= \frac{1}{2}\langle\mathbb{E}[(\bar{\eta}_k\bar{g}_k)^{\otimes 2} \mid \mathcal{F}_k], \nabla^2 q(u_k)\rangle \\
&\quad - \frac{n\bar{\eta}_k^2}{2}\nu_c(u_k)\mathcal{P}(u_k)\frac{1}{n}\operatorname{tr}(\Sigma\nabla^2 q(u_k)) \\
\Delta\mathcal{M}_k^{\mathsf{Noise}}(q) &:= 2c\sqrt{d}\sigma_k\langle b_k, \nabla q(u_k)\rangle + \frac{1}{2\sqrt{n}}\langle 2c\sqrt{\frac{d}{n}}\sigma_k b_k n\bar{\eta}_k\bar{g}_k^\top, \nabla^2 q(u_k)\rangle \\
&\quad + \frac{d}{2}\langle(2c\sigma_k b_k)^{\otimes 2}, \nabla^2 q(u_k)\rangle - 2dc^2\langle\sigma_k^2 I_d, \nabla^2 q(u_k)\rangle \\
\mathbb{E}[\Delta\mathcal{E}_k^{\mathsf{Noise}}(q) \mid \mathcal{F}_k] &:= 2dc^2\langle\left(\sigma_k^2 - \frac{1}{n^3}\tilde{\sigma}(k/n)^2\right) I_d, \nabla^2 q(u_k)\rangle, \\
\mathbb{E}[\Delta\mathcal{E}_k^{\mathsf{Step}}(q) \mid \mathcal{F}_k] &:= \left(\frac{\bar{\eta}(k/d)}{n} - \bar{\eta}_k\right)\mu_c(u_k)\langle\Sigma u_k, \nabla q(u_k)\rangle \\
&\quad + \left(\frac{\bar{\eta}_k^2}{2} - \frac{\bar{\eta}(k/n)^2}{2n^2}\right)\nu_c(u_k)\mathcal{P}(u_k)\operatorname{tr}(\Sigma\nabla^2 q(u_k)).
\end{aligned} \tag{J.41}$$

We note that the first three terms are in common with the analysis in Marshall et al. [2025], while the last three are the result of the private noise and the adaptive learning rate step in Algorithm 9.1.

In a similar way, we introduce the shorthand $V_t = \Theta_t - \theta^*$ such that $V_0 = u_0$. Using Itô's formula on (9.22), for any quadratic function q, we have that

$$\begin{aligned}
\mathrm{d}q(V_t) = &-\bar{\eta}(t)\mu_c(V_t)\langle\Sigma V_t, \nabla q(V_t)\rangle\mathrm{d}t + \bar{\eta}^2(t)\nu_c(V_t)\mathcal{P}(V_t)\frac{1}{n}\operatorname{tr}\left(\Sigma\nabla^2 q(V_t)\right)\mathrm{d}t \\
&+ 2\frac{d}{n^2}c^2\tilde{\sigma}^2(t)\operatorname{tr}\left(\nabla^2 q(V_t)\right)\mathrm{d}t + \mathrm{d}\mathcal{M}_t^{\mathsf{H\text{-}DP\text{-}GD}},
\end{aligned} \tag{J.42}$$

where we introduced the shorthand

$$\mathrm{d}\mathcal{M}_t^{\mathsf{H\text{-}DP\text{-}GD}} := \langle\nabla q(V_t), \sqrt{\frac{2\bar{\eta}(t)^2\nu_c(\Theta_t)\mathcal{P}(\Theta_t)\Sigma}{n} + 4\frac{d}{n^2}c^2\tilde{\sigma}(t)^2 I_d}\,\mathrm{d}B_t\rangle, \tag{J.43}$$

with B_t being a d-dimensional standard Brownian motion.

Let M be a positive constant that will be fixed later. The dynamic is first controlled up to the stopping time

$$\tau := \inf\{k : \|u_k\|_2 \geq M \cup \lfloor tn\rfloor : \|V_t\|_2 \geq M\}. \tag{J.44}$$

Then, we will denote the stopped processes $u_k^\tau = u_{k\wedge\tau}$ and $V_t^\tau = V_{t\wedge(\tau/n)}$, which will be the objects we will compare in the following arguments. This stopping time is introduced for technical reasons, and we will later show that $\tau \geq n$.

Taking the difference between (J.40) and (J.42), and following the same argument in Lemma 1 in Marshall et al. [2025], we get that there exists an absolute constants $C_1 = C_1(\|\Sigma\|_{\mathrm{op}}, c, \bar{\eta})$,

such that

$$\sup_{0\le t<1}\left|q(u^\tau_{\lfloor tn\rfloor})-q(V^\tau_t)\right|\le C_1\int_0^1\sup_{q\in Q}\left|q(u^\tau_{\lfloor sn\rfloor})-q(V^\tau_s)\right|\mathrm{d}s$$
$$+\sup_{0\le t<(1\wedge(\tau/n))}\left(|\sum_{k=1}^{\lfloor tn\rfloor}\mathbb{E}[\Delta\mathcal{E}_k^{\mathsf{Hess}}(q)\mid\mathcal{F}_k]|+|\sum_{k=1}^{\lfloor tn\rfloor}\mathbb{E}[\Delta\mathcal{E}_k^{\mathsf{Noise}}(q)\mid\mathcal{F}_k]|+|\sum_{k=1}^{\lfloor tn\rfloor}\mathbb{E}[\Delta\mathcal{E}_k^{\mathsf{Step}}(q)\mid\mathcal{F}_k]|\right)$$
$$+\sup_{0\le t<(1\wedge(\tau/n))}\left(|\mathcal{M}^{\mathsf{Grad}}_{\lfloor tn\rfloor}(q)|+|\mathcal{M}^{\mathsf{Hess}}_{\lfloor tn\rfloor}(q)|+|\mathcal{M}^{\mathsf{Noise}}_{\lfloor tn\rfloor}(q)|+|\mathcal{M}^{\mathsf{H\text{-}DP\text{-}GD}}_t(q)|\right)+O(d^{-1}),\tag{J.45}$$

where we introduced the notation $\mathcal{M}_k(q)=\sum_{j=1}^k\Delta\mathcal{M}_j(q)$ and $\mathcal{M}^{\mathsf{H\text{-}DP\text{-}GD}}_t(q)=\int_0^t\mathrm{d}\mathcal{M}^{\mathsf{H\text{-}DP\text{-}GD}}_s(q)$. The last term follows from transitioning (J.40) to continuous times, which involves an additional discretization error of $O(d^{-1})$ (see also Section A.3 in Collins-Woodfin et al. [2024a] for more details). Importantly, differently from Lemma 1 in Marshall et al. [2025], we miss a term proportional to $\left(\mathcal{R}(\Theta_s)+\mathcal{R}(\theta_{\lfloor sn\rfloor})\right)^{-1/2}$ in (J.45). This is possible since, as $\zeta>0$, both $\mu_c(\theta)$ and $\nu_c(\theta)\mathcal{P}(\theta)$ are Lipschitz with respect to $\mathcal{R}(\theta)$ with constant-order Lipschitz constant due to Lemma J.5, strengthening the condition in their Assumption 5.

Denoting with $\mathcal{M}$ the sum of the last two lines in (J.45), by Lemma 2 in Marshall et al. [2025] (for $|\mathcal{M}^{\mathsf{Grad}}_{\lfloor tn\rfloor}(q)|$, $|\mathcal{M}^{\mathsf{Hess}}_{\lfloor tn\rfloor}(q)|$ and $|\sum_{k=1}^{\lfloor tn\rfloor}\mathbb{E}[\Delta\mathcal{E}_k^{\mathsf{Hess}}(q)\mid\mathcal{F}_k]|$), Lemma J.8 (for $|\mathcal{M}^{\mathsf{Noise}}_{\lfloor tn\rfloor}(q)|$ and $|\sum_{k=1}^{\lfloor tn\rfloor}\mathbb{E}[\Delta\mathcal{E}_k^{\mathsf{Noise}}(q)\mid\mathcal{F}_k]|$), Lemma J.9 (for $|\mathcal{M}^{\mathsf{H\text{-}DP\text{-}GD}}_t(q)|$), and Lemma J.11 (for $|\mathcal{M}^{\mathsf{Step}}_{\lfloor tn\rfloor}(q)|$), there are two constants $C_2(\|\Sigma\|_{\mathrm{op}},\bar\eta,M,c,\gamma)$ and $C_3(\|\Sigma\|_{\mathrm{op}},\bar\eta,M,c)>0$ such that

$$\mathcal{M}\le C_2n^{-1/2}\left(\log^2n+C_3\right),\tag{J.46}$$

with probability at least $1-e^{-\log^2n}$.

Then, following Lemma 3 in Marshall et al. [2025], we can define a set $\bar Q\subseteq Q$ with $|\bar Q|\le C_4(\|\Sigma\|_{\mathrm{op}})d^4$, such that, for all $q\in Q$, there exists a $\bar q\in\bar Q$ that satisfies $\|q-\bar q\|_{C^2}\le d^{-2}$. Then, taking the union bound over this set yields

$$\sup_{q\in Q}\sup_{0\le t<1}\left|q(u^\tau_{\lfloor tn\rfloor})-q(V^\tau_t)\right|\le C_1\int_0^1\sup_{q\in Q}\left|q(u^\tau_{\lfloor sn\rfloor})-q(V^\tau_s)\right|\mathrm{d}s+\mathcal{M}+C_5d^{-2},\tag{J.47}$$

with probability at least $1-C_4d^4e^{-\log^2n}$. Thus, with this same probability, the application of Gronwall's inequality gives

$$\sup_{q\in Q}\sup_{0\le t<1}\left|q(u^\tau_{\lfloor tn\rfloor})-q(V^\tau_t)\right|\le(\mathcal{M}+C_5d^{-2})\exp(C_1)=O\left(\frac{\log^2n}{\sqrt{n}}\right),\tag{J.48}$$

where the last step hides a dependence on M (see (J.44)). Then, we are left to show that $\tau\ge n$ for M chosen to be a constant independent of d and n. This follows the same approach as in Lemma 4 in Marshall et al. [2025], which is here formalized in Lemma J.10. In particular, in Lemma J.10, we show that there exists a constant $C_6(\|\Sigma\|_{\mathsf{op}},c,\bar\eta)>0$, such that for any $r\ge0$, with probability at least $1-2e^{-r^2/2}$, it holds that $\sup_{t\in[0,1]}\|u_{\lfloor tn\rfloor}\|^2\le\|u_0\|^2e^{C_6(1+d^{-1/2}r)}$ and $\sup_{t\in[0,1]}\|V_t\|^2\le\|V_0\|^2e^{C_6(1+d^{-1/2}r)}$. Thus, M can be chosen as a constant independent from d and n, such that $\tau>n$ with overwhelming probability, and the thesis follows. □

Proof of Remark 9.7. The result follows the same proof of Theorem J.7, but considering $q(\theta-\theta^*)=\left\|\Sigma^{1/2}(\theta-\theta^*)\right\|_2^2/2$. While this function is not included in Q, it is still such that $\|q\|_{C^2}=O(1)$, allowing the same argument to go through. Alternatively, it can be recovered by the Cauchy integral formula taking $\frac{i}{4\pi}\oint_\Omega zq(\cdot)\mathrm{d}z$ (see, e.g., (J.77)). □

J.3.1 Technical lemmas for Theorem J.7

In this section, we present the statements and proofs of the technical lemmas used in the proof of Theorem J.7. All notation is defined as needed throughout, and Assumptions 9.3 and 9.4 are assumed to hold unless otherwise specified. For notational simplicity, we use the shorthand $\|\cdot\|$ to denote the Frobenius norm for matrices.

Lemma J.8. *For any quadratic $q \in Q$ such that $\|q\|_{C^2} \leq 1$, it holds that*

$$\left|\sum_{k=1}^{\lfloor tn \rfloor} \mathbb{E}[\Delta \mathcal{E}_k^{\mathsf{Noise}}(q) \mid \mathcal{F}_k]\right| \leq \frac{2\gamma_n}{\rho^2 n} C_{\eta,2}, \quad \textit{a.s.} \tag{J.49}$$

where $C_{\eta,2}$ denotes the upper bound on the absolute value of the second derivative of $\tilde{\eta}^2(t)$. In addition, there is a constant $C = C(c, \gamma, \rho, M) > 0$, such that, for any $y \in [1, n]$,

$$\sup_{1 \leq k \leq (n \wedge \tau)} |\mathcal{M}_k^{\mathsf{Noise}}(q)| \leq C n^{-\frac{1}{2}} y, \tag{J.50}$$

with probability at least $1 - e^{-y}$.

Proof. Recall the definition in (J.41)

$$\begin{aligned} \Delta \mathcal{M}_k^{\mathsf{Noise}}(q) = 2c\sqrt{d}\sigma_k \langle b_k, \nabla q(u_k) \rangle + \frac{1}{2\sqrt{n}} 2c \sqrt{\frac{d}{n}} \sigma_k n \bar{\eta}_k \bar{g}_k^\top \nabla^2 q(u_k) b_k \\ + 2dc^2\sigma_k^2 \operatorname{tr}((b_k)^{\otimes 2} \nabla^2 q(u_k)) - 2dc^2\sigma_k^2 \operatorname{tr}(\nabla^2 q(u_k)). \end{aligned} \tag{J.51}$$

Then, for any $k \leq \tau$, we rewrite the martingale as a combination of the following three terms

$$\Delta \mathcal{M}_k^{\mathsf{Noise}}(q) = \frac{1}{n^{3/2}} \langle b_k, A_{k,1} + A_{k,2} \rangle + \frac{1}{n^2} \operatorname{tr}\left(b_k^{\otimes 2} - I_d\right), \tag{J.52}$$

where we introduced the shorthands

$$A_{k,1} := 2cn^{3/2}\sqrt{d}\sigma_k \nabla q(u_k), \; A_{k,2} := cn\sqrt{\frac{d}{n}} \sigma_k n \bar{\eta}_k \nabla^2 q(u_k) \bar{g}_k, \; C_k := 2dn^2 c^2 \sigma_k^2 \nabla^2 q(u_k). \tag{J.53}$$

We will separately bound the contribution of each term in terms of its Orlicz norm. Let us start with the second term, and consider $q \in Q$. This is a quadratic function of the iterates; its Hessian, therefore, does not depend on the iterates and explicitly on k. In addition, we have that $\sup_{z \in \Omega} \|R(z; \Sigma)\|_{\mathsf{op}} \leq 2$. Since $b_k \sim \mathcal{N}(0, I_d)$ and independent, using the Hanson-Wright inequality (Theorem 6.2.1 in Vershynin [2018]) we have that, for some $c > 0$ and $K = \max_{i \in [d]} \|b_{k+1,i}\|_{\psi_2}$,

$$\begin{aligned} \mathbb{P}\left(\frac{1}{n^2} \operatorname{tr}\left((b_k^{\otimes 2} - I_d) C_k\right) \geq t\right) &\leq 2 \exp\left(-c \min\left(\frac{t^2 n^4}{K^4 \|C_k\|^2}, \frac{tn^2}{K^2 \|C_k\|_{\mathsf{op}}}\right)\right) \\ &\leq 2 \exp\left(-c' \min\left(\frac{t^2 n^4}{d}, tn^2\right)\right), \end{aligned}$$

for some $c' > 0$ independent of k, d, n. To justify the last passage, note that, by the structure of the noise,

$$\sigma_k^2 \leq \frac{1}{\rho^2 n^2} |\tilde{\eta}(k/n)^2 - \tilde{\eta}((k+1)/n)^2| \leq \frac{2}{\rho^2 n^3} \max_{x \in [\frac{k}{n}, \frac{k+1}{n}]} \left|\frac{\mathsf{d}}{\mathsf{d}x} \tilde{\eta}^2(x)\right| \leq \frac{2}{\rho^2 n^3} C_{\eta,1},$$

as $\left|\frac{\mathsf{d}}{\mathsf{d}x}\tilde{\eta}^2(x)\right| \leq C_{\eta,1}$ for some $C_{\eta,1}$ due to Assumption 9.4. Using the above bound, we obtain that $\|C_k\|_{\mathrm{op}} \leq 2dn^2c^2\sigma_k^2\|q\|_{C^2} \leq 8c^2/\rho^2 C_{\eta,1}$, and similarly $\|C_k\| \leq 8\sqrt{d}c^2/\rho^2 C_{\eta,1}$. We, therefore, have that $\|\frac{1}{n^2}\langle b_k^{\otimes 2} - I_d, C_k\rangle\|_{\psi_1} \leq Cn^{-1}$, for some constant $C(\rho, c, C_{\eta,1}) > 0$.

For the first term, by Eq. (89) in Marshall et al. [2025], the norm of q is bounded for the stopped process u_k^τ as follows:

$$\|\nabla q(u)\| \leq \|\nabla^2 q\|_{\mathsf{op}}\|u\| + \|\nabla q(0)\| \leq \|q\|_{C^2}(1 + \|u\|) \leq C(1 + M). \tag{J.54}$$

Hence, we have that $\|A_{k,1}\| \leq 2n^{3/2}c\sqrt{n}\sigma_k C(1+M) \leq 4\sqrt{d}c\sqrt{C_{\eta,1}/\gamma_n}C(1+M)/\rho$, which implies that

$$\left\|\frac{1}{n^{3/2}}\langle b_k, A_{k,1}\rangle\right\|_{\psi_2} \leq \frac{C}{n}, \tag{J.55}$$

with $C = C(M, \rho, \gamma_n, C_{\eta,1}, c) > 0$. Then, we have

$$\begin{aligned}\left|\frac{1}{n^{3/2}}\langle b_k, A_{k,2}\rangle\right| &\leq \frac{1}{n^{3/2}}cn\sqrt{\gamma_n}\sigma_k\tilde{\eta}(k/n)\left|\langle b_k, \nabla^2 q(u_k)g_k\rangle\right| \qquad \text{(J.56)}\\ &\leq \frac{2c\sqrt{C_{\eta,1}}}{n\sqrt{\gamma_n}\rho}\left|\langle b_k, \nabla^2 q(u_k)x_{k+1}\rangle(\langle x_{k+1}, u_k\rangle + z_{k+1})\right|.\end{aligned}$$

Due to (J.54) and Assumptions 9.3 and 9.4, we have $\|\langle b_k, \nabla^2 q(u_k)x_{k+1}\rangle\|_{\psi_1} \leq C\|b_k\|_{\psi_2}\|x_{k+1}\|_{\psi_2} \leq C$ for some $C > 0$ that depends on $\|\Sigma\|_{\mathrm{op}}$. Finally, we note that $\|\langle x_{k+1}, u_k\rangle\|_{\psi_2} \leq C(1+M)$, and $\|z_k\|_{\psi_2} \leq C\zeta$ for any $k \leq \tau$. Combining the above, we obtain that there is a constant $C = C(c, C_{\eta,1}, \rho, \gamma_n, M, \zeta) > 0$ such that

$$\phi_{k,1} := \inf\{t > 0 : \mathbb{E}[\exp(|\Delta\mathcal{M}_k^{\mathsf{Noise}}(q)|/t) \mid \mathcal{F}_{k-1}] \leq 2\} \leq Cn^{-1}. \tag{J.57}$$

We then apply Lemma 5 in Marshall et al. [2025] for some absolute constant $C > 0$ for all $t > 0$:

$$\begin{aligned}\mathbb{P}\left(\sup_{1\leq k\leq n\wedge\tau}|\mathcal{M}_k^{\mathsf{Noise}}(q) - \mathbb{E}\mathcal{M}_0^{\mathsf{Noise}}(q)| \geq t\right) &\leq 2\exp\left(-\min\{\frac{t}{C\max_{k\in[n]}\phi_{k,1}}, \frac{t^2}{C\sum_{i=1}^n\phi_{i,1}^2}\}\right) \qquad \text{(J.58)}\\ &\leq 2\exp\left(-Cn\min\{t, t^2\}\right).\end{aligned}$$

As we assume that n is proportional to d and noting that $\mathbb{E}\mathcal{M}_0^{\mathsf{Noise}}(q) = 0$ by definition, we then have that, for any $y \in [1, n]$,

$$\sup_{1\leq k\leq(n\wedge\tau)}|\mathcal{M}_k^{\mathsf{Noise}}(q)| \leq Cn^{-\frac{1}{2}}y, \tag{J.59}$$

with probability at least $1 - e^{-y}$ for any $y \in [1, n]$.

Next, we bound the error due to the discretization:

$$\begin{aligned}\left|\sum_{k=1}^{\lfloor tn\rfloor}\mathbb{E}[\Delta\mathcal{E}_k^{\mathsf{Noise}}(q) \mid \mathcal{F}_k]\right| &\leq 2dc^2\sum_{k=1}^{\lfloor tn\rfloor}\left|\sigma_k^2 - \frac{1}{n^3}\tilde{\sigma}(k/n)^2\right|\cdot\left|\mathrm{tr}(\nabla^2 q(u_k))\right|\\ &\leq \frac{2d^2\|q\|_{C^2}}{n^2\rho^2}c^2\sum_{k=1}^{\lfloor tn\rfloor}\left|\tilde{\eta}(k/n)^2 - \tilde{\eta}((k+1)/n)^2 - \frac{1}{n}\tilde{\sigma}(k/n)^2\right|\\ &\leq \frac{2\gamma_n^2c^2\|q\|_{C^2}}{\rho^2n^2}\sum_{k=1}^{\lfloor tn\rfloor}\max_{x\in(\frac{k}{n},\frac{k+1}{n})}\left|\frac{\mathsf{d}^2}{\mathsf{d}x^2}\tilde{\eta}(x)^2\right|,\end{aligned}$$

where we use the definition of the noise function as the derivative of the learning rate $\rho^2\tilde{\sigma}(x)^2 = -\frac{\mathrm{d}}{\mathrm{d}x}\tilde{\eta}^2(x)$ at any point $x \in [0,1)$. Then, as $|\frac{\mathrm{d}^2}{\mathrm{d}x^2}\tilde{\eta}^2(x)| \leq C_{\eta,2}$ for some constant $C_{\eta,2}$, the desired result follows. □

Lemma J.9. *Denote by $\mathcal{M}_t^{\text{H-DP-GD},\tau}$ the H-DP-GD martingale in which the stopping time is imposed. There is a constant $C = C(c, C_{\eta,1}, \gamma_n, \|\Sigma\|_{\mathrm{op}}, M) > 0$, such that for any quadratic $q \in Q$ with $\|q\|_{C^2} \leq 1$, and for any $y \in [1, n]$, we have*

$$\sup_{0\leq t<1} |\mathcal{M}_t^{\text{H-DP-GD},\tau}(q)| \leq Cn^{-\frac{1}{2}}y, \tag{J.60}$$

with probability at least $1 - e^{-y}$.

Proof. To bound the martingale error from H-DP-GD under some general statistic $q \in Q$, differently from the argument to obtain Eq. (72) in Lemma 2 in Marshall et al. [2025], we need to control the additional term due to the additive noise in Algorithm 9.1.

Using Itô's formula for any quadratic function q:

$$\begin{aligned}
\mathrm{d}q(V_t) =& -\bar{\eta}(t)\mu_c(\Theta_t)\nabla\mathcal{P}(\Theta_t)^\top\nabla q(V_t)\mathrm{d}t + \bar{\eta}^2(t)\nu_c(\Theta_t)\mathcal{P}(\Theta_t)\frac{1}{n}\operatorname{tr}\left(\Sigma\nabla^2 q\right)\mathrm{d}t \\
&+ 2\frac{d}{n^2}c^2\tilde{\sigma}^2(t)\operatorname{tr}\left(\nabla^2 q\right)\mathrm{d}t + \mathrm{d}\mathcal{M}_t^{\text{H-DP-GD}}, \\
\mathrm{d}\mathcal{M}_t^{\text{H-DP-GD}} :=& \langle\nabla q(V_t), \sqrt{\frac{2\bar{\eta}(t)^2\nu_c(\Theta_t)\mathcal{P}(\Theta_t)\Sigma}{n} + 4\frac{d}{n^2}c^2\tilde{\sigma}(t)^2 I_d}\,\mathrm{d}B_t\rangle,
\end{aligned} \tag{J.61}$$

with B_t being a standard d dimensional Brownian motion. The quadratic variation of the martingale is then bounded a.s.

$$\langle\mathcal{M}(q)\rangle_t \leq \frac{Ct}{n} + \frac{d}{n^2}(4c^2\bar{\eta}^2\|\Sigma\|_{\mathrm{op}}(1+M)^2),$$

for some constant $C = C(\|\Sigma\|_{\mathrm{op}}, M, C_{\eta,1}, c, \gamma_n) > 0$, where we used Assumption 9.4 which gives $|\rho^2\tilde{\sigma}(t)| = |\frac{\mathrm{d}}{\mathrm{d}t}\tilde{\eta}^2(t)| \leq C_{\eta,1}$. The claim is then proved by an application of Gaussian concentration inequalities as in Section B.6 in Marshall et al. [2025]. □

Lemma J.10. *For any $r \geq 0$, with probability at least $1 - 2e^{-r^2/2}$, we jointly have*

$$\sup_{t\in[0,1)} \|V_t\|^2 \leq \|V_0\|^2 e^{C(1+d^{-1/2}r)}, \qquad \sup_{t\in[0,1)} \|u_{\lfloor nt\rfloor}\|^2 \leq \|u_0\|^2 e^{C(1+d^{-1/2}r)}. \tag{J.62}$$

Proof. The proof follows a path similar to the one of Lemma 4 in Marshall et al. [2025]. In particular, consider the function $\varphi(V_t) = \log(1 + \|V_t\|^2)$. Then, by application of Itô's Lemma to (9.22),

$$\begin{aligned}
\mathrm{d}\varphi(V_t) =& -\bar{\eta}(t)\frac{\mu_c(\Theta_t)}{1+\|V_t\|^2}\nabla\mathcal{P}(\Theta_t)^\top V_t\mathrm{d}t \\
&+ 2\bar{\eta}^2(t)\frac{\nu_c(\Theta_t)\mathcal{P}(\Theta_t)}{(1+\|V_t\|^2)^2}\frac{1}{n}\operatorname{tr}\left(\Sigma(I_d(1+\|V_t\|^2) - V_t\otimes V_t)\right)\mathrm{d}t \\
&+ 2\frac{d}{n}c^2\tilde{\sigma}^2(t)\frac{1}{(1+\|V_t\|^2)^2}\mathrm{d}t + \mathrm{d}\mathcal{M}_t(\varphi),
\end{aligned} \tag{J.63}$$

with

$$\mathrm{d}\mathcal{M}_t(\varphi) = \frac{1}{(1+\|V_t\|^2)}\langle V_t, \sqrt{\frac{2\bar\eta(t)^2\nu_c(\Theta_t)\mathcal{P}(\Theta_t)\Sigma}{n} + 4\frac{d}{n^2}c^2\tilde\sigma(t)^2 I_d}\,\mathrm{d}B_t\rangle. \tag{J.64}$$

The drift terms and the quadratic variation terms can be bounded by some $C(c, \|\Sigma\|_{\mathrm{op}}, \zeta, \gamma_n, M)$. The quadratic variation of the martingale term is bounded as $\langle M\rangle_t \le \frac{Ct}{n}$. We then have, by Gaussian concentration,

$$\mathbb{P}\left(\sup_{0\le t<1}\varphi(V_t) \ge C(1 + r/\sqrt{n})\right) \le e^{-r^2/2}, \tag{J.65}$$

which proves the first part of the claim. The bound on $\sup_{t\in[0,1)}\|u_{\lfloor nt\rfloor}\|^2$ follows the same strategy as the argument following Lemma 4 in Marshall et al. [2025]. □

Lemma J.11. *For any quadratic $q \in Q$ such that $\|q\|_{C^2} \le 1$,*

$$\sup_{1\le k\le(n\wedge\tau)} |\sum_{k=1}^{n}\mathbb{E}[\Delta\mathcal{E}_k^{\mathsf{Step}}(q) \mid \mathcal{F}_k]| \le Cn^{-\frac{1}{2}}\log^2 n, \tag{J.66}$$

with probability at least $1 - e^{-c\log^2 d}$.

Proof. Recall that

$$\begin{aligned}\mathbb{E}[\Delta\mathcal{E}_k^{\mathsf{Step}}(q) \mid \mathcal{F}_k] := &\left(\frac{\bar\eta(k/d)}{n} - \bar\eta_k\right)\mu_c(u_k)\langle\Sigma u_k, \nabla q(u_k)\rangle \\ &+ \left(\frac{\bar\eta_k^2}{2} - \frac{\bar\eta(k/n)^2}{2n^2}\right)\nu_c(u_k)\mathcal{P}(u_k)\operatorname{tr}(\Sigma\nabla^2 q(u_k)),\end{aligned} \tag{J.67}$$

and $\|u_k\|_2^2 < M$ by the definition of τ. Thus, this implies

$$|\mu_c(u_k)\langle\Sigma u_k, \nabla q(u_k)\rangle| = O(1), \qquad \left|\nu_c(u_k)\mathcal{P}(u_k)\operatorname{tr}(\Sigma\nabla^2 q(u_k))\right| = O(d). \tag{J.68}$$

First, notice that

$$\frac{\bar\eta(k/d)}{n} - \bar\eta_k = \min\left(\frac{\tilde\eta(k/d)}{n}, \frac{2}{d}\right) - \min\left(\frac{\tilde\eta(k/d)}{n}, \frac{2}{\|x_k\|_2^2}\right), \tag{J.69}$$

which implies

$$\left|\frac{\bar\eta(k/d)}{n} - \bar\eta_k\right| \le 2\left|\frac{1}{d} - \frac{1}{\|x_k\|_2^2}\right| = \frac{2\left|\|x_k\|_2^2 - d\right|}{d\,\|x_k\|_2^2}. \tag{J.70}$$

Similarly, it can be shown that

$$\frac{1}{2}\left|\frac{\bar\eta(k/d)^2}{n^2} - \bar\eta_k^2\right| \le \frac{2\left|\|x_k\|_2^2 - d\right|\left(\|x_k\|_2^2 + d\right)}{d^2\,\|x_k\|_2^4}. \tag{J.71}$$

Hanson-Wright inequality gives

$$\mathbb{P}\left(\left|\|x_k\|_2^2 - d\right| > \sqrt{d}\log d\right) \le 2\exp\left(-c_1\log^2 d\right), \tag{J.72}$$

which is sufficient to show that, with probability $1 - 2\exp\left(-c_1\log^2 d\right)$,

$$\left|\mathbb{E}[\Delta\mathcal{E}_k^{\mathsf{Step}}(q) \mid \mathcal{F}_k]\right| = O\left(\frac{\log d}{d^{3/2}}\right), \tag{J.73}$$

which yields the thesis after performing a union bound over $k \in [n]$. □

J.3.2 Proof of Theorem 9.6

Consider (J.61), which reads

$$\begin{aligned}\mathsf{d}q(V_t) =& -\bar{\eta}(t)\mu_c(\Theta_t)\nabla\mathcal{P}(\Theta_t)^\top\nabla q(V_t)\mathsf{d}t + \bar{\eta}^2(t)\nu_c(\Theta_t)\mathcal{P}(\Theta_t)\frac{1}{n}\operatorname{tr}\left(\Sigma\nabla^2 q\right)\mathsf{d}t \\ &+ 2\frac{d}{n^2}c^2\tilde{\sigma}^2(t)\operatorname{tr}\left(\nabla^2 q\right)\mathsf{d}t + \mathsf{d}\mathcal{M}_t^{\text{H-DP-GD}}(q),\end{aligned} \tag{J.74}$$

where we recall the notation $V_t = \Theta_t - \theta^*$, and $q(V_t) = V_t^\top(\Sigma - zI)^{-1}V_t/2$, and where in Lemma J.9 we showed that $\sup_{0\le t<1}|\int_0^t \mathcal{M}_\tau^{\text{H-DP-GD}}(q)\mathsf{d}\tau| \le Cn^{-\frac{1}{2}}y$ with probability at least $1-e^{-y}$. Then, we have

$$\begin{aligned}\mathsf{d}q(V_t) =& -\bar{\eta}(t)\mu_c(\Theta_t)V_t^\top\Sigma(\Sigma - zI)^{-1}V_t\mathsf{d}t + \bar{\eta}^2(t)\nu_c(\Theta_t)\mathcal{P}(\Theta_t)\frac{1}{n}\operatorname{tr}\left(\Sigma(\Sigma-zI)^{-1}\right)\mathsf{d}t \\ &+ 2\frac{d}{n^2}c^2\tilde{\sigma}^2(t)\operatorname{tr}\left((\Sigma-zI)^{-1}\right)\mathsf{d}t + \mathsf{d}\mathcal{M}_t^{\text{H-DP-GD}}(q),\end{aligned} \tag{J.75}$$

which, using the identity $\Sigma(\Sigma - zI)^{-1} = I + z(\Sigma - zI)^{-1}$, can be written as

$$\begin{aligned}\mathsf{d}q(V_t) =& -\bar{\eta}(t)\mu_c(\Theta_t)\left(\|V_t\|_2^2 + zq(V_t)\right)\mathsf{d}t + \bar{\eta}^2(t)\nu_c(\Theta_t)\mathcal{P}(\Theta_t)\gamma_n\frac{\operatorname{tr}\left(\Sigma(\Sigma-zI)^{-1}\right)}{d}\mathsf{d}t \\ &+ 2c^2\tilde{\sigma}^2(t)\gamma_n^2\frac{\operatorname{tr}\left((\Sigma-zI)^{-1}\right)}{d}\mathsf{d}t + \mathsf{d}\mathcal{M}_t^{\text{H-DP-GD}}(q).\end{aligned} \tag{J.76}$$

Notice that

$$q(V_t) = \frac{1}{2}\sum_{i=1}^d \frac{\langle V_t, \omega_i\rangle^2}{\lambda_i - z}, \qquad \|V_t\|_2^2 = \frac{i}{\pi}\oint_\Omega q(V_t), \qquad \mathcal{R}(\Theta_t) = \frac{i}{2\pi}\oint_\Omega zq(V_t), \tag{J.77}$$

where the last two equations follow from Cauchy integral formula. Then, we can define the following (deterministic) integro-differential equation after removing the martingale term from (J.61):

$$\begin{aligned}\mathsf{d}\hat{q}(z,t) =& -\bar{\eta}(t)\mu_c\left(\frac{i}{2\pi}\oint_\Omega z\hat{q}(z,t)\right)\left(\frac{i}{\pi}\oint_\Omega \hat{q}(z,t) + z\hat{q}(z,t)\right)\mathsf{d}t \\ &+ \bar{\eta}^2(t)\nu_c\left(\frac{i}{2\pi}\oint_\Omega z\hat{q}(z,t)\right)\left(\frac{\zeta^2}{2} + \frac{i}{2\pi}\oint_\Omega z\hat{q}(z,t)\right)\gamma_n\frac{\operatorname{tr}\left(\Sigma(\Sigma-zI)^{-1}\right)}{d}\mathsf{d}t \\ &+ 2c^2\tilde{\sigma}^2(t)\gamma_n^2\frac{\operatorname{tr}\left((\Sigma-zI)^{-1}\right)}{d}\mathsf{d}t,\end{aligned} \tag{J.78}$$

where all differentials are taken with respect to time. (J.78) is defined for $z \in \Omega$, $t \in [0,1]$, and is such that $\hat{q}(z,0) = q(V_0) = \theta^{*\top}(\Sigma - zI)^{-1}\theta^*/2$.

Using the same argument as in the proof of Theorem J.7 based on Gronwall's inequality, relying on $\zeta > 0$ which makes both $\mu_c(\theta)$ and $\nu_c(\theta)\mathcal{P}(\theta)$ Lipschitz with respect to $\mathcal{R}(\theta)$ with constant-order Lipschitz constant, and since the length of Ω is bounded, we have

$$\sup_{t\in[0,1)}|\mathcal{R}(\Theta_t) - R(t)| = O\left(\frac{\log^2 n}{\sqrt{n}}\right), \tag{J.79}$$

with overwhelming probability, where we have introduce the shorthand $R(t) = \frac{i}{4\pi}\oint_\Omega z\hat{q}(z,t)$.

Then, due to Lemma 4.1 in Collins-Woodfin et al. [2024a], we have that the unique solution of (J.78) is

$$\hat{q}(z,t) = \frac{1}{d}\sum_{i=1}^{d}\frac{D_i(t)}{\lambda_i - z}, \tag{J.80}$$

where $D_i(t)$ is defined according to (9.15). Multiplying both sides by z, integrating it over Ω and using Cauchy formula, together with Theorem J.7, gives the first desired result, i.e.,

$$\sup_{t\in[0,1)}\left|\mathcal{R}(\theta_{\lfloor tn\rfloor}) - R(t)\right| = O\left(\frac{\log^2 n}{\sqrt{n}}\right), \tag{J.81}$$

with overwhelming probability.

To prove the same inequality in expectation, define the event $\mathcal{E}$ such that

$$\Delta := \sup_{t\in[0,1)}\left|\mathcal{R}(\theta_{\lfloor tn\rfloor}) - R(t)\right| \geq C_1\frac{\log^2 n}{\sqrt{n}}. \tag{J.82}$$

By (J.81), there exists a constant C_1 such that $\mathbb{P}(\mathcal{E}) \leq 2e^{-c_1\log^2 d}$. Then

$$\sup_{t\in[0,1)}\left|\mathbb{E}\left[\mathcal{R}(\theta_{\lfloor tn\rfloor})\right] - R(t)\right| \leq \sup_{t\in[0,1)}\mathbb{E}\left[\left|\mathcal{R}(\theta_{\lfloor tn\rfloor}) - R(t)\right|\right] \leq C_1\frac{\log^2 n}{\sqrt{n}} + \mathbb{E}\left[\Delta\mathbf{1}_{\mathcal{E}}\right]. \tag{J.83}$$

Notice that

$$\Delta \leq \sup_{t\in[0,1)}\left|\mathcal{R}(\theta_{\lfloor tn\rfloor})\right| + C_2 \leq C_3\left(1 + \sup_{t\in[0,1)}\left\|u_{\lfloor nt\rfloor}\right\|_2^2\right), \tag{J.84}$$

since $R(t)$ exists in $t\in[0,1]$ and does not depend on d and n, and since $\|\Sigma\|_{\mathsf{op}} = O(1)$. Thus, Cauchy-Schwartz inequality yields

$$\mathbb{E}\left[\Delta\mathbf{1}_{\mathcal{E}}\right]^2 \leq \mathbb{E}\left[\Delta^2\right]\mathbb{P}(\mathcal{E}) \leq 2C_4e^{-c_1\log^2 d}\left(1 + \mathbb{E}\left[\sup_{t\in[0,1)}\left\|u_{\lfloor nt\rfloor}\right\|_2^4\right]\right). \tag{J.85}$$

By Lemma J.10, we have that for any $r\geq 0$, $Z = \sup_{t\in[0,1]}\left\|u_{\lfloor nt\rfloor}\right\|_2^2 \leq \|u_0\|^2 e^{C_4(1+d^{-1/2}r)}$ with probability at least $1-2e^{-r^2/2}$. This bound implies that $2\log Z$ has a dimension free sub-Gaussian tail, which in turn implies that $\mathbb{E}[e^{2\log Z}] = \mathbb{E}\left[\sup_{t\in[0,1)}\left\|u_{\lfloor nt\rfloor}\right\|_2^4\right] = O(1)$, giving

$$\mathbb{E}\left[\Delta\mathbf{1}_{\mathcal{E}}\right]^2 \leq C_5e^{-c_1\log^2 d} = O\left(\frac{1}{n}\right). \tag{J.86}$$

Merging in (J.83) gives the desired result. □

J.3.3 Proof of Proposition 9.8

Due to the update rule in Algorithm 9.1, we have

$$\begin{aligned}
2\mathcal{R}(\theta_n) &= \left\|\Sigma^{1/2}\left(\theta_n - \theta^*\right)\right\|_2^2 \\
&= \left\|\Sigma^{1/2}\left(\theta_{n-1} - \bar{\eta}_n\bar{g}_n + 2C_{\mathsf{clip}}\sigma_n b_n - \theta^*\right)\right\|_2^2 \\
&= \left\|\Sigma^{1/2}\left(\theta_{n-1} - \theta^* + 2C_{\mathsf{clip}}\sigma_n b_n\right)\right\|_2^2 + \left\|\Sigma^{1/2}\bar{\eta}_n\bar{g}_n\right\|_2^2 \\
&\quad - 2\bar{\eta}_n\bar{g}_n^\top\Sigma\left(\theta_{n-1} - \theta^*\right) - 4\bar{\eta}_n\bar{g}_n^\top\Sigma C_{\mathsf{clip}}\sigma_n b_n.
\end{aligned} \tag{J.87}$$

By Assumptions 9.3 and 9.4, and due to the definition of $\bar{g}_n$, we have

$$\left\|\Sigma^{1/2}\bar{\eta}_n\bar{g}_n\right\|_2 \leq \|\Sigma\|_{\mathsf{op}}^{1/2}\,|\bar{\eta}_n|\,\|\bar{g}_n\|_2 = O\left(\frac{\sqrt{d}}{n}\right) = O\left(\frac{1}{\sqrt{d}}\right). \tag{J.88}$$

Theorem 3.1.1 in Vershynin [2018] also guarantees that $\|b_n\|_2 = O(\sqrt{d}+u)$ with probability at least $1-2\exp\left(-c_1u^2\right)$, for some absolute constant $c_1>0$. Thus, with this probability, we have

$$\left|4\bar{\eta}_n\bar{g}_n^\top\Sigma C_{\mathsf{clip}}\sigma_n b_n\right| = \left|4\bar{\eta}_n\bar{g}_n^\top\Sigma C_{\mathsf{clip}}\frac{\eta_n}{\rho}b_n\right| = O\left(\frac{1}{n}\sqrt{d}\sqrt{d}\,\frac{1}{n}(\sqrt{d}+u)\right) = O\left(\frac{1}{\sqrt{d}}+\frac{u}{d}\right), \tag{J.89}$$

which gives

$$\left|2\mathcal{R}(\theta_n) - \left\|\Sigma^{1/2}\left(\theta_{n-1}-\theta^*+2C_{\mathsf{clip}}\sigma_n\Sigma^{1/2}b_n\right)\right\|_2^2\right| = O\left(\frac{1+\sqrt{\mathcal{R}(\theta_{n-1})}}{\sqrt{d}}+\frac{u}{d}\right). \tag{J.90}$$

Similarly, we have

$$\begin{aligned}\left\|\Sigma^{1/2}\left(\theta_{n-1}-\theta^*+2C_{\mathsf{clip}}\sigma_n b_n\right)\right\|_2^2 &= \left\|\Sigma^{1/2}\left(\theta_{n-1}-\theta^*\right)\right\|_2^2 + \left\|2C_{\mathsf{clip}}\sigma_n\Sigma^{1/2}b_n\right\|_2^2 \\ &\quad +4C_{\mathsf{clip}}\sigma_n b_n^\top\Sigma\left(\theta_{n-1}-\theta^*\right).\end{aligned} \tag{J.91}$$

However, since b_n is a standard Gaussian vector, we have that, with probability at least $1-2\exp\left(-c_2u^2\right)$,

$$\begin{aligned}\left|4C_{\mathsf{clip}}\sigma_n b_n^\top\Sigma\left(\theta_{n-1}-\theta^*\right)\right| &\leq 4C_{\mathsf{clip}}\sigma_n\left\|\Sigma^{1/2}\right\|_{\mathsf{op}}\left\|\Sigma^{1/2}\left(\theta_{n-1}-\theta^*\right)\right\|_2 u \\ &= 4c\sqrt{d}\frac{\tilde{\eta}(1)}{\rho n}\left\|\Sigma^{1/2}\right\|_{\mathsf{op}}\left\|\Sigma^{1/2}\left(\theta_{n-1}-\theta^*\right)\right\|_2 u \\ &= O\left(\frac{\sqrt{d}u}{n}\right)\left\|\Sigma^{1/2}\left(\theta_{n-1}-\theta^*\right)\right\|_2 \\ &= O\left(\frac{\sqrt{\mathcal{R}(\theta_{n-1})}u}{\sqrt{d}}\right).\end{aligned} \tag{J.92}$$

Then, an application of the Hanson-Wright inequality (see Theorem 6.2.1 in Vershynin [2018]) yields

$$\left|\left\|2C_{\mathsf{clip}}\sigma_n\Sigma^{1/2}b_n\right\|_2^2 - 4c^2d\frac{\tilde{\eta}^2(1)}{\rho^2n^2}\operatorname{tr}(\Sigma)\right| \leq 4c^2d\frac{\tilde{\eta}^2(1)}{\rho^2n^2}\|\Sigma\|u = O\left(\frac{u}{\sqrt{d}}\right), \tag{J.93}$$

with probability at least $1-2\exp\left(-c_3\min(\sqrt{d}u,u^2)\right)$.

Putting everything together gives

$$\left|\mathcal{R}(\theta_n)-\mathcal{R}(\theta_{n-1})-2c^2\tilde{\eta}^2(1)\gamma_n^2/\rho^2\right| = O\left(\frac{\left(1+\sqrt{\mathcal{R}(\theta_{n-1})}\right)(u+1)}{\sqrt{d}}\right), \tag{J.94}$$

with probability at least $1-2\exp\left(-c_4\min(\sqrt{d}u,u^2)\right)$.

Setting $u = \log d$, we have that

$$\left|\mathcal{R}(\theta_n) - \mathcal{R}(\theta_{n-1}) - 2c^2\tilde{\eta}^2(1)\gamma_n^2/\rho^2\right| = O\left(\frac{\log d}{\sqrt{d}}\right), \tag{J.95}$$

with overwhelming probability, where we used $\mathrm{tr}(\Sigma) = d$ and Theorem 9.6, which implies $\mathcal{R}(\theta_{n-1}) = O(1)$ with overwhelming probability. Then, the first part of Theorem 9.6 and the triangle inequality concludes the first part of the proof.

For the result in expectation, notice that we have

$$\begin{aligned}\left|\mathbb{E}\left[\mathcal{R}(\theta_n)\right] - R(1) - 2c^2\tilde{\eta}^2(1)\gamma_n^2/\rho^2\right| \leq \mathbb{E}\left[\left|\mathcal{R}(\theta_n) - \mathcal{R}(\theta_{n-1}) - 2c^2\tilde{\eta}^2(1)\gamma_n^2/\rho^2\right|\right] \\ + \left|\mathbb{E}\left[\mathcal{R}(\theta_{n-1})\right] - R(1)\right|.\end{aligned} \tag{J.96}$$

The second term on the RHS is $O(\log^2 n/\sqrt{n})$ because of the second part of the statement of Theorem 9.6, and due to continuity of $R(t)$ in $t = 1$. The argument of the first term of the RHS is the product of two random variables. The first one is such that

$$\mathbb{E}\left[\left(1 + \sqrt{\mathcal{R}(\theta_{n-1})}\right)^2\right] \leq 2 + 2\mathbb{E}\left[\mathcal{R}(\theta_{-1})\right] = O(1), \tag{J.97}$$

due to the second part of the statement of Theorem 9.6. The second one, due to (J.94), is a random variable Z such that $Z \leq (1+u)/\sqrt{d}$ with probability at least $1-2\exp\left(-c_4\min(\sqrt{d}u, u^2)\right)$. Thus, $\sqrt{d}Z$ is sub-exponential with norm smaller than a constant, which implies

$$\mathbb{E}\left[Z^2\right] = O\left(\frac{1}{d}\right). \tag{J.98}$$

Plugging (J.97) and (J.98) in (J.96), and using Cauchy-Schwartz inequality, gives the desired result. □

J.4 Proofs for Section 9.5

In this section, all asymptotic notations are with respect to the limit $\gamma \to 0$.

Proposition J.12. *Define $\overline{R}(t), \underline{R}(t) : [0,1] \to \mathbb{R}$ as the unique solutions of the following ODEs*

$$\begin{aligned}\mathrm{d}\overline{R}(t) &= -2\lambda_{\min}\tilde{\eta}(t)\mu_c(\overline{R})\overline{R}\mathrm{d}t + \lambda_{\max}\tilde{\eta}^2(t)\nu_c(\overline{R})(\overline{R} + \zeta^2/2)\gamma\mathrm{d}t + 2c^2\tilde{\sigma}^2(t)\gamma^2\mathrm{d}t, \\ \mathrm{d}\underline{R}(t) &= -2\lambda_{\max}\tilde{\eta}(t)\mu_c(\underline{R})\underline{R}\mathrm{d}t + \tilde{\eta}^2(t)\nu_c(\underline{R})(\underline{R} + \zeta^2/2)\gamma\mathrm{d}t + 2c^2\tilde{\sigma}^2(t)\gamma^2\mathrm{d}t,\end{aligned} \tag{J.99}$$

where $\overline{R}(0) = \underline{R}(0) = R(0) = \|\Sigma^{1/2}\theta^*\|_2^2/2$. Then, for every $t \in [0,1]$, we have

$$\underline{R}(t) \leq R(t) \leq \overline{R}(t), \tag{J.100}$$

where $R(t)$ is defined in Definition 9.5 after taking $\gamma_n \to \gamma$.

Proof. Until the end of the proof, we will define more auxiliary ODEs, such that the RHS is uniformly Lipschitz in the dependent variable at all times $t \in [0,1]$. Then, by the extension of

the Picard–Lindelöf theorem (see Corollary 2.6 in Teschl [2012]), we have that their solutions exist and are unique, and therefore also have continuous derivatives. Then, defining

$$\begin{aligned}
\mathsf{d}\overline{D}_i &= -2\lambda_{\min}\tilde{\eta}(t)\mu_c(R(t))\overline{D}_i\mathsf{d}t + \lambda_i\tilde{\eta}^2(t)\nu_c(R(t))(R(t)+\zeta^2/2)\gamma\mathsf{d}t + 2c^2\tilde{\sigma}^2(t)\gamma^2\mathsf{d}t,\\
\mathsf{d}\underline{D}_i &= -2\lambda_{\max}\tilde{\eta}(t)\mu_c(R(t))\underline{D}_i\mathsf{d}t + \lambda_i\tilde{\eta}^2(t)\nu_c(R(t))(R(t)+\zeta^2/2)\gamma\mathsf{d}t + 2c^2\tilde{\sigma}^2(t)\gamma^2\mathsf{d}t,
\end{aligned} \tag{J.101}$$

by standard ODE comparison arguments (see Theorem 1.3 in Teschl [2012]) we have that $\overline{D}_i(t) \geq D_i(t) \geq \underline{D}_i(t)$ for all $t \in [0,1]$. Then, averaging over i (weighting by λ_i) the equations in (J.101) and (9.15), and defining $\overline{R}'(t) = \frac{1}{d}\sum_{i=1}^d \lambda_i \overline{D}_i(t)$ and $\underline{R}'(t) = \frac{1}{d}\sum_{i=1}^d \lambda_i \underline{D}_i(t)$, we get $\overline{R}'(t) \geq R(t) \geq \underline{R}'(t)$ for all $t \in [0,1]$, with

$$\begin{aligned}
\mathsf{d}\overline{R}' &= -2\lambda_{\min}\tilde{\eta}(t)\mu_c(R(t))\overline{R}'\mathsf{d}t + \frac{\operatorname{tr}(\Sigma^2)}{d}\tilde{\eta}^2(t)\nu_c(R(t))(R(t)+\zeta^2/2)\gamma\mathsf{d}t + 2c^2\tilde{\sigma}^2(t)\gamma^2\mathsf{d}t,\\
\mathsf{d}\underline{R}' &= -2\lambda_{\max}\tilde{\eta}(t)\mu_c(R(t))\underline{R}'\mathsf{d}t + \frac{\operatorname{tr}(\Sigma^2)}{d}\tilde{\eta}^2(t)\nu_c(R(t))(R(t)+\zeta^2/2)\gamma\mathsf{d}t + 2c^2\tilde{\sigma}^2(t)\gamma^2\mathsf{d}t.
\end{aligned} \tag{J.102}$$

Furthermore, for a fixed value of c, we have that the functions $\mu_c(R)$ and $\nu_c(R)(R+\zeta^2/2)$ are respectively monotonically decreasing and increasing with respect to R (see Lemma J.4). Then, since by Jensen inequality we have that $\operatorname{tr}(\Sigma^2) \geq \operatorname{tr}(\Sigma) = d$ (where the last step holds due to Assumption 9.3), and since we also have $\operatorname{tr}(\Sigma^2) \geq \lambda_{\max}\operatorname{tr}(\Sigma) = \lambda_{\max} d$, by defining

$$\begin{aligned}
\mathsf{d}\overline{R} &= -2\lambda_{\min}\tilde{\eta}(t)\mu_c(\overline{R})\overline{R}\mathsf{d}t + \lambda_{\max}\tilde{\eta}^2(t)\nu_c(\overline{R})(\overline{R}+\zeta^2/2)\gamma\mathsf{d}t + 2c^2\tilde{\sigma}^2(t)\gamma^2\mathsf{d}t,\\
\mathsf{d}\underline{R} &= -2\lambda_{\max}\tilde{\eta}(t)\mu_c(\underline{R})\underline{R}\mathsf{d}t + \tilde{\eta}^2(t)\nu_c(\underline{R})(\underline{R}+\zeta^2/2)\gamma\mathsf{d}t + 2c^2\tilde{\sigma}^2(t)\gamma^2\mathsf{d}t,
\end{aligned} \tag{J.103}$$

again due to Theorem 1.3 in Teschl [2012] we get that $\overline{R}(t) \geq \overline{R}'(t) \geq R(t) \geq \underline{R}'(t) \geq \underline{R}(t)$ for all $t \in [0,1]$, which gives the desired result. □

J.4.1 Proofs on output perturbation and constant private noise

All the ODEs defined in this (and the following) section will be such that their RHS is uniformly Lipschitz in the dependent variable at all times $t \in [0,1]$, which in turn guarantees they have a unique solution. Furthermore, the RHSs will also be uniformly Lipschitz with respect to the variable t due to Assumption 9.4. Thus, if $R(0) = R'(0)$, and

$$\mathsf{d}R = f(t,R), \qquad \mathsf{d}R' = f'(t,R'), \tag{J.104}$$

with

$$f'(t,R'(t)) \geq f(t,R'(t)), \tag{J.105}$$

for all $t \in [0,1]$, we have that

$$R'(t) \geq R(t) \text{ for all } t \in [0,1]. \tag{J.106}$$

The same statement holds also for the opposite inequality, and will be used extensively to bound the solutions of $R(t)$ for different schedules. This is a direct application of Theorem 1.3 in Teschl [2012].

To ease the presentation, we introduce the notation $v = \tilde{\eta}(0)$. We will provide the proof separately for $\alpha = 0$ and $\alpha = 1/2$, and the proof for $\alpha \geq 1$ in the next section. We will also denote the test risk at initialization $R(0) = \left\|\Sigma^{1/2}\theta^*\right\|_2^2/2$, and all asymptotic notations will be with respect to the limit $\gamma \to 0$.

The proof of Proposition 9.9 is given separately for the case $\alpha = 0$ and $\alpha = 1/2$, in the general case where $\lambda_{\max}$ and $\lambda_{\min}$ are allowed to depend on γ. The results are stated separately in Propositions J.14 and J.15, and Proposition 9.9 is a direct consequence of them. At the end of this section we will discuss the sub-optimality of $c = \omega(1)$, to prove (9.44).

Lemma J.13. *Let $\beta < 1$ and $E_\beta(x) : \mathbb{R}^+ \to \mathbb{R}^+$ be the exponential integral function defined as*

$$E_\beta(x) := \int_1^{+\infty} \frac{e^{-xt}}{t^\beta} \, \mathrm{d}t. \tag{J.107}$$

Then, we have that

$$\lim_{x \to 0^+} x^{1-\beta} E_\beta(x) = \Gamma(1-\beta), \tag{J.108}$$

where $\Gamma(\cdot)$ denotes the Euler Gamma function

$$\Gamma(s) := \int_0^{\infty} e^{-t} t^{s-1} \, \mathrm{d}t. \tag{J.109}$$

Furthermore, we have that, for all $x \geq 0$,

$$E_\beta(x) \leq \Gamma(1-\beta) x^{-1+\beta}. \tag{J.110}$$

Finally, if either $\beta \geq 0$ or $x \geq -2\beta$, we also have that

$$E_\beta(x) \leq \frac{2e^{-x}}{x}. \tag{J.111}$$

Proof. The change of variable $u = xt$ in the definition of $E_\beta(x)$ yields

$$\begin{aligned} E_\beta(x) &= x^{\beta-1} \int_x^{+\infty} \frac{e^{-u}}{u^\beta} \, \mathrm{d}u \\ &= x^{\beta-1} \left(\int_0^{+\infty} e^{-u} u^{-\beta} \, \mathrm{d}u - \int_0^x e^{-u} u^{-\beta} \, \mathrm{d}u \right) \\ &= x^{\beta-1} \left(\Gamma(1-\beta) - \int_0^x e^{-u} u^{-\beta} \, \mathrm{d}u \right). \end{aligned} \tag{J.112}$$

Then, (J.108) and (J.110) readily follow from the fact that the last term in the equation above is bounded by

$$0 \leq \int_0^x e^{-u} u^{-\beta} \, \mathrm{d}u \leq \int_0^x u^{-\beta} \, \mathrm{d}u = \frac{x^{1-\beta}}{1-\beta}. \tag{J.113}$$

For the upper bound, denoting with

$$\Gamma(s, x) := \int_x^{\infty} e^{-t} t^{s-1} \, \mathrm{d}t \tag{J.114}$$

the upper incomplete Euler gamma function, (J.112) allows us to write

$$E_\beta(x) = \frac{1}{x^{1-\beta}} \Gamma(1-\beta, x). \tag{J.115}$$

Via an integration by parts, we have

$$\Gamma(1-\beta, x) = \int_x^{\infty} e^{-t} t^{-\beta} \, \mathrm{d}t = e^{-x} x^{-\beta} - \beta \int_x^{\infty} e^{-t} t^{-\beta-1} \, \mathrm{d}t. \tag{J.116}$$

If $\beta \geq 0$, we have $\Gamma(1-\beta, x) \leq e^{-x}x^{-\beta}$, which together with (J.115) gives (J.111). If $\beta < 0$, the second term in the equation above is positive, and since $t \geq x$, it can be upper bounded as

$$-\beta \int_x^\infty e^{-t}t^{-\beta-1}\,\mathrm{d}t \leq -\frac{\beta}{x}\int_x^\infty e^{-t}t^{-\beta}\,\mathrm{d}t = -\frac{\beta}{x}\,\Gamma(1-\beta, x), \tag{J.117}$$

which, if plugged in (J.116), for $x \geq -2\beta$ gives

$$\Gamma(1-\beta, x) \leq \frac{e^{-x}x^{-\beta}}{1+\beta/x} \leq 2e^{-x}x^{-\beta}, \tag{J.118}$$

and the thesis again follows from (J.115). □

$\alpha = 0$: output perturbation. Recall that in the setting $\alpha = 0$, we have

$$\begin{aligned}
\mathrm{d}\overline{R}(t) &= -2\lambda_{\min} vc\frac{\mu_c(\overline{R})}{c}\overline{R}\mathrm{d}t + \lambda_{\max}(vc)^2\frac{\nu_c(\overline{R})(\overline{R}+\zeta^2/2)}{c^2}\gamma\mathrm{d}t,\\
\mathrm{d}\underline{R}(t) &= -2\lambda_{\max} vc\frac{\mu_c(\underline{R})}{c}\underline{R}\mathrm{d}t + (vc)^2\frac{\nu_c(\underline{R})(\underline{R}+\zeta^2/2)}{c^2}\gamma\mathrm{d}t.
\end{aligned} \tag{J.119}$$

Importantly, recall that the risk $\mathcal{R}(\theta^p)$ in this setting is not well approximated by $R(1)$, due to Proposition 9.8.

Proposition J.14. *Let Assumptions 9.3 and 9.4 hold, and let θ^p be the solution obtained with Algorithm 9.1, with the schedule defined in* (9.33) *for $\alpha = 0$, in the setting $\gamma = d/n \to 0$. Furthermore, assume*

$$\frac{\lambda_{\max}}{\lambda_{\min}^2}\ln(1/\gamma)\gamma = o(1). \tag{J.120}$$

Then, by setting

$$c = O(1), \qquad vc = \frac{C\ln(1/\gamma)}{\lambda_{\min}}, \qquad v \leq 2/\gamma, \tag{J.121}$$

for a large enough constant C which does not depend on γ, ρ, Σ, we have that, with overwhelming probability,

$$\mathcal{R}(\theta^p) = O\left(\frac{\lambda_{\max}\gamma\ln(1/\gamma)}{\lambda_{\min}^2} + \frac{\gamma^2\ln^2(1/\gamma)}{\rho^2\lambda_{\min}^2}\right). \tag{J.122}$$

Furthermore, suppose there exists $h > 0$ such that $\rho = \Omega\left(\gamma^{1-h}\right)$. Then, for any choice of the hyper-parameters c and v such that $v \leq 2/\gamma$, we have that

$$\mathcal{R}(\theta^p) = \Omega\left(\frac{\gamma\ln(1/\gamma)}{\lambda_{\max}^2} + \frac{\gamma^2\ln^2(1/\gamma)}{\rho^2\lambda_{\max}^2}\right). \tag{J.123}$$

Proof. Let us introduce the shorthands

$$\begin{aligned}
\overline{f}(t,\overline{R}) &= -2\lambda_{\min} vc\frac{\mu_c(\overline{R})}{c}\overline{R}\mathrm{d}t + \lambda_{\max}(vc)^2\frac{\nu_c(\overline{R})(\overline{R}+\zeta^2/2)}{c^2}\gamma\mathrm{d}t,\\
\underline{f}(t,\underline{R}) &= -2\lambda_{\max} vc\frac{\mu_c(\underline{R})}{c}\underline{R}\mathrm{d}t + (vc)^2\frac{\nu_c(\underline{R})(\underline{R}+\zeta^2/2)}{c^2}\gamma\mathrm{d}t,
\end{aligned} \tag{J.124}$$

corresponding to the RHSs of the ODEs of interest. Then, consider the auxiliary ODEs

$$\mathrm{d}\overline{R}' = -\overline{a}\overline{R}'\mathrm{d}t + \overline{b}\mathrm{d}t =: \overline{f}'(t,\overline{R}')\mathrm{d}t, \qquad \mathrm{d}\underline{R}' = -\underline{a}\,\underline{R}'\mathrm{d}t + \underline{b}\mathrm{d}t =: \underline{f}'(t,\underline{R}')\mathrm{d}t, \qquad \overline{a},\overline{b},\underline{a},\underline{b} > 0, \tag{J.125}$$

with initial conditions $\overline{R}'(0) = \overline{R}(0) = R(0) = \underline{R}(0) = \underline{R}'(0)$.

Notice that $\overline{R}'(t)$ admits the closed form solution

$$\overline{R}'(t) = \left(R(0) - \overline{b}/\overline{a}\right) e^{-\overline{a}t} + \overline{b}/\overline{a}. \tag{J.126}$$

Similarly, we have that

$$\underline{R}'(t) = \left(R(0) - \underline{b}/\underline{a}\right) e^{-\underline{a}t} + \underline{b}/\underline{a}. \tag{J.127}$$

Since $c = O(1)$, by Lemma J.5, we have that

$$\frac{\underline{c}_\mu(c,\zeta)}{\sqrt{2R+\zeta^2}} < \frac{\mu_c(R)}{c} < \frac{1}{\sqrt{\pi\left(R+\zeta^2/2\right)}}, \qquad \underline{c}_\nu(c,\zeta) < \frac{2\nu_c(R)\left(R+\zeta^2/2\right)}{c^2} < 1. \tag{J.128}$$

Then, let us set

$$\overline{a} = 2\lambda_{\min} vc \frac{\underline{c}_\mu(c,\zeta)}{\sqrt{R(0)+1+\zeta^2}}, \quad \overline{b} = \lambda_{\max}\frac{v^2c^2\gamma}{2}, \quad \underline{a} = 2vc\frac{\lambda_{\max}}{\sqrt{\pi\zeta^2/2}}, \quad \underline{b} = \frac{v^2c^2\gamma\underline{c}_\nu(c,\zeta)}{2}. \tag{J.129}$$

As $cv = C\ln(1/\gamma)/\lambda_{\min}$, the choice in (J.129) ensures that $\underline{b}/\underline{a} \le 1$ and $\overline{b}/\overline{a} \le 1$ as long as $\lambda_{\max}/\lambda_{\min}^2 \ln(1/\gamma)\gamma = o(1)$, which in turn guarantees that

$$\overline{R}'(t) \in [0, R(0)+1], \qquad \underline{R}'(t) \in [0, R(0)+1]. \tag{J.130}$$

Thus, (J.129) guarantees

$$\overline{f}(t,\overline{R}'(t)) < \overline{f}'(t,\overline{R}'(t)), \qquad \underline{f}(t,\underline{R}'(t)) > \underline{f}'(t,\underline{R}'(t)), \qquad \text{for all } \ t\in[0,1]. \tag{J.131}$$

Thus, by (J.106), we have

$$\underline{R}'(t) < \underline{R}(t), \qquad \overline{R}'(t) > \overline{R}(t) \qquad \text{for all } \ t\in(0,1]. \tag{J.132}$$

In particular, plugging $vc = C\ln(1/\gamma)/\lambda_{\min}$ and (J.129) in (J.126), we have that

$$\overline{R}(1) < \overline{R}'(1) < e^{-\overline{a}} + \frac{\overline{b}}{\overline{a}} = O\left(\frac{\lambda_{\max}}{\lambda_{\min}^2}\ln(1/\gamma)\gamma\right), \tag{J.133}$$

as long as C is chosen to be large enough. Then, the upper bounds follows from Theorem 9.6 and Propositions 9.8 and J.12, as the risk at the last iterate roughly increases by $2c^2v^2\gamma^2/\rho^2 = O(\gamma^2\ln^2(1/\gamma)/(\rho^2\lambda_{\min}^2))$.

For the lower bound, let us first suppose $c \le 1$, which implies that (J.128) holds with $\underline{c}_\mu(c,\zeta)$ and $\underline{c}_\nu(c,\zeta)$ lower bounded by constants independent of γ. Pick $\underline{a}, \underline{b}$ as in (J.129).

First, let us consider the case $\underline{a} \le 1$. We have that (J.132) gives

$$\underline{R}(1) > \underline{R}'(1) = R(0)e^{-\underline{a}} + \frac{\underline{b}}{\underline{a}} - \frac{\underline{b}}{\underline{a}}e^{-\underline{a}} > R(0)e^{-\underline{a}} \ge R(0)/e = \Omega(1). \tag{J.134}$$

In the case $\underline{a} > 1$, (J.132) gives

$$\underline{R}(1) > \underline{R}'(1) = R(0)e^{-\underline{a}} + \frac{\underline{b}}{\underline{a}} - \frac{\underline{b}}{\underline{a}}e^{-\underline{a}} > R(0)e^{-\underline{a}} + \frac{\underline{b}}{\underline{a}}(1-e^{-1}) = R(0)\left(e^{-vc\lambda_{\max}a} + \frac{vcb\gamma}{\lambda_{\max}a}\right), \tag{J.135}$$

where a, b are positive constants which do not depend on γ, v, c and the spectrum of Σ. Then, let us consider the following quantity

$$\underline{\mathcal{R}} := R(0)\left(e^{-vc\lambda_{\max}a} + \frac{vcb\gamma}{\lambda_{\max}a}\right) + 2v^2c^2\frac{\gamma^2}{\rho^2}. \tag{J.136}$$

We have that

$$\arg\min_{vc} e^{-vc\lambda_{\max}a} + \frac{vcb\gamma}{\lambda_{\max}a} = \frac{1}{\lambda_{\max}a}\ln\left(\frac{(\lambda_{\max}a)^2}{b\gamma}\right), \tag{J.137}$$

which implies

$$\underline{\mathcal{R}} > R(0)\left(e^{-vc\lambda_{\max}a} + \frac{vcb\gamma}{\lambda_{\max}a}\right) = \Omega\left(\frac{\gamma\ln(1/\gamma)}{\lambda_{\max}^2}\right). \tag{J.138}$$

Note that, assuming $\rho = \Omega(\gamma^{1-h})$, we have

$$\begin{aligned}
&\min_{vc\geq\frac{\ln(\rho^2/\gamma^2)}{2a\lambda_{\max}}} e^{-vc\lambda_{\max}a} + v^2c^2\frac{\gamma^2}{\rho^2} \geq \min_{vc\geq\frac{\ln(\rho^2/\gamma^2)}{2a\lambda_{\max}}} v^2c^2\frac{\gamma^2}{\rho^2} = \Omega\left(\frac{\gamma^2\ln^2(1/\gamma)}{\rho^2\lambda_{\max}^2}\right),\\
&\min_{0\leq vc\leq\frac{\ln(\rho^2/\gamma^2)}{2a\lambda_{\max}}} e^{-vc\lambda_{\max}a} + v^2c^2\frac{\gamma^2}{\rho^2} \geq \ \frac{\gamma}{\rho} = \Omega\left(\frac{\gamma^2\ln^2(1/\gamma)}{\rho^2}\right) = \Omega\left(\frac{\gamma^2\ln^2(1/\gamma)}{\rho^2\lambda_{\max}^2}\right).
\end{aligned} \tag{J.139}$$

Then, merging (J.136), (J.138) and (J.139) yields

$$\underline{\mathcal{R}} = \Omega\left(\frac{\gamma\ln(1/\gamma)}{\lambda_{\max}^2} + \frac{\gamma^2\ln^2(1/\gamma)}{\rho^2\lambda_{\max}^2}\right). \tag{J.140}$$

The result in (J.135) together with Theorem 9.6 and Propositions 9.8 and J.12 imply that, with overwhelming probability,

$$\mathcal{R}(\theta^p) = \Omega\left(\frac{\gamma\ln(1/\gamma)}{\lambda_{\max}^2} + \frac{\gamma^2\ln^2(1/\gamma)}{\rho^2\lambda_{\max}^2}\right). \tag{J.141}$$

It remains to prove the lower bound for $c > 1$. From (J.10), one readily has that $0 < \mu_c(R) < 1$. Note that

$$\nu_c(R)(R + \zeta^2/2) \geq \max\left(\frac{\nu_c(R)(R+\zeta^2/2)}{c^2}, \nu_c(R)\zeta^2/2\right), \tag{J.142}$$

which combined with (J.26) gives that $\nu_c(R)(R + \zeta^2/2) > b_1$, for a positive constant b_1 independent of γ, v, c and the spectrum of Σ. Furthermore, (J.24) implies that $\nu_c(R)(R + \zeta^2/2) < c^2/2$. Thus, the solution of $\underline{R}$ is lower bounded by that of the ODE below:

$$\mathsf{d}\underline{R}'' = -2\lambda_{\max}v\underline{R}''\mathsf{d}t + v^2b_1\gamma\mathsf{d}t, \tag{J.143}$$

with initial condition $\underline{R}''(0) = R(0)$. Thus, following the same steps we used to achieve (J.140), it can be shown that, as $c \geq 1$, we have, with overwhelming probability,

$$\mathcal{R}(\theta^p) = \Omega\left(\frac{\gamma\ln(1/\gamma)}{\lambda_{\max}^2} + \frac{\gamma^2\ln^2(1/\gamma)}{\rho^2\lambda_{\max}^2}\right), \tag{J.144}$$

which gives the desired result. □

$\alpha = 1/2$**: constant private noise.** In the setting $\alpha = 1/2$, (9.33) gives

$$\begin{aligned}
\mathrm{d}\overline{R}(t) &= -2\lambda_{\min} vc\sqrt{1-t}\frac{\mu_c(\overline{R})}{c}\overline{R}\mathrm{d}t + \lambda_{\max}(vc)^2(1-t)\frac{\nu_c(\overline{R})(\overline{R}+\zeta^2/2)}{c^2}\gamma\mathrm{d}t + 2(vc)^2\frac{\gamma^2}{\rho^2}\mathrm{d}t,\\
\mathrm{d}\underline{R}(t) &= -2\lambda_{\max} vc\sqrt{1-t}\frac{\mu_c(\underline{R})}{c}\underline{R}\mathrm{d}t + (vc)^2(1-t)\frac{\nu_c(\underline{R})(\underline{R}+\zeta^2/2)}{c^2}\gamma\mathrm{d}t + 2(vc)^2\frac{\gamma^2}{\rho^2}\mathrm{d}t.
\end{aligned} \tag{J.145}$$

Importantly, recall that the risk $\mathcal{R}(\theta^p)$ in this setting is well approximated by $R(1)$, due to Proposition 9.8, since $\tilde{\eta}(1) = 0$.

Proposition J.15. *Let Assumptions 9.3 and 9.4 hold, and let θ^p be the solution obtained with Algorithm 9.1, with the schedule defined in* (9.33) *for $\alpha = 1/2$, in the setting $\gamma = d/n \to 0$. Furthermore, assume*

$$\frac{\ln^2(1/\gamma)\gamma}{\lambda_{\min}^2}\left(\lambda_{\max} + \frac{\gamma}{\rho^2}\right) = o(1). \tag{J.146}$$

Then, by setting

$$c = O(1), \qquad vc = \frac{C\ln(1/\gamma)}{\lambda_{\min}}, \qquad v \le 2/\gamma, \tag{J.147}$$

for a large enough constant C which does not depend on γ, ρ, Σ, we have that, with overwhelming probability,

$$\mathcal{R}(\theta^p) = O\left(\frac{\lambda_{\max}\gamma\ln^{2/3}(1/\gamma)}{\lambda_{\min}^2} + \frac{\gamma^2\ln^{4/3}(1/\gamma)}{\rho^2\lambda_{\min}^2}\right). \tag{J.148}$$

Furthermore, suppose there exists $h > 0$ such that $\rho = \Omega\left(\gamma^{1-h}\right)$. Then, for any choice of the hyper-parameters c and v such that $v \le 2/\gamma$, we have that

$$\mathcal{R}(\theta^p) = \Omega\left(\frac{\gamma\ln^{2/3}(1/\gamma)}{\lambda_{\max}^2} + \frac{\gamma^2\ln^{4/3}(1/\gamma)}{\rho^2\lambda_{\max}^2}\right). \tag{J.149}$$

Proof. As before, let us introduce the shorthands

$$\begin{aligned}
\overline{f}(t,\overline{R}) &= -2\lambda_{\min} vc\sqrt{1-t}\frac{\mu_c(\overline{R})}{c}\overline{R} + \lambda_{\max}(vc)^2(1-t)\frac{\nu_c(\overline{R})(\overline{R}+\zeta^2/2)}{c^2}\gamma + 2(vc)^2\frac{\gamma^2}{\rho^2},\\
\underline{f}(t,\underline{R}) &= -2\lambda_{\max} vc\sqrt{1-t}\frac{\mu_c(\underline{R})}{c}\underline{R} + (vc)^2(1-t)\frac{\nu_c(\underline{R})(\underline{R}+\zeta^2/2)}{c^2}\gamma + 2(vc)^2\frac{\gamma^2}{\rho^2}.
\end{aligned} \tag{J.150}$$

Then, consider the auxiliary ODEs

$$\begin{aligned}
\mathrm{d}\overline{R}' &= -\overline{a}\sqrt{1-t}\overline{R}'\mathrm{d}t + \overline{b}_1(1-t)\mathrm{d}t + \overline{b}_2\mathrm{d}t =: \overline{f}'(t,\overline{R})\mathrm{d}t, \qquad \overline{a},\overline{b}_1,\overline{b}_2 > 0,\\
\mathrm{d}\underline{R}' &= -\underline{a}\sqrt{1-t}\underline{R}'\mathrm{d}t + \underline{b}_1(1-t)\mathrm{d}t + \underline{b}_2\mathrm{d}t =: \underline{f}'(t,\underline{R})\mathrm{d}t, \qquad \underline{a},\underline{b}_1,\underline{b}_2 > 0,
\end{aligned} \tag{J.151}$$

with initial conditions $\overline{R}'(0) = \overline{R}(0) = R(0) = \underline{R}(0) = \underline{R}'(0)$, which admit the closed-form solutions

$$\begin{aligned}
\overline{R}'(t) = \frac{R(0)}{3}e^{-2\overline{a}(1-(1-t)^{3/2})/3}\Bigg(3 + \\
- \frac{2}{R(0)}\overline{b}_1 e^{2\overline{a}/3}\left(E_{-1/3}\left(\frac{2\overline{a}}{3}\right) - (1-t)^2 E_{-1/3}\left(\frac{2\overline{a}(1-t)^{3/2}}{3}\right)\right)\\
- \frac{2}{R(0)}\overline{b}_2 e^{2\overline{a}/3}\left(E_{1/3}\left(\frac{2\overline{a}}{3}\right) - (1-t) E_{1/3}\left(\frac{2\overline{a}(1-t)^{3/2}}{3}\right)\right)\Bigg),
\end{aligned} \tag{J.152}$$

$$\underline{R}'(t) = \frac{R(0)}{3} e^{-2\underline{a}(1-(1-t)^{3/2})/3} \Bigg(3 + \\ - \frac{2}{R(0)} \underline{b}_1 e^{2\underline{a}/3} \left(E_{-1/3}\left(\frac{2\underline{a}}{3}\right) - (1-t)^2 E_{-1/3}\left(\frac{2\underline{a}(1-t)^{3/2}}{3}\right)\right) \\ - \frac{2}{R(0)} \underline{b}_2 e^{2\underline{a}/3} \left(E_{1/3}\left(\frac{2\underline{a}}{3}\right) - (1-t) E_{1/3}\left(\frac{2\underline{a}(1-t)^{3/2}}{3}\right)\right)\Bigg), \tag{J.153}$$

expressed in terms of the exponential integral functions $E_{-1/3}(\cdot)$, $E_{1/3}(\cdot)$ defined in (J.107).

Note that, for $t \in [0,1]$, $\overline{R}'(t) \geq 0$. In fact, if this is not the case, by continuity of $\overline{R}'$, there exists an interval $(t^*, t^* + \delta) \subseteq [0,1]$ s.t. $\overline{R}'(t^*) = 0$ and $\overline{R}'(t) < 0$ for all $t \in (t^*, t^* + \delta)$. However, if $\overline{R}'(t^*) = 0$, then the derivative of $\overline{R}'$ evaluated at t^* is $\geq b_2 > 0$ by (J.151), which is a contradiction. A similar argument gives that, for $t \in [0,1]$, $\underline{R}'(t) \geq 0$.

Next, we upper bound $\overline{R}'(t)$ as

$$\overline{R}'(t) \leq R(0) + \frac{2}{3}\overline{b}_1 e^{2\overline{a}(1-t)^{3/2}/3}(1-t)^2 E_{-1/3}\left(\frac{2\overline{a}(1-t)^{3/2}}{3}\right) \\ + \frac{2}{3}\overline{b}_2 e^{2\overline{a}(1-t)^{3/2}/3}(1-t) E_{1/3}\left(\frac{2\overline{a}(1-t)^{3/2}}{3}\right). \tag{J.154}$$

If $2\overline{a}(1-t)^{3/2}/3 \geq 2/3$, an application of (J.111) gives that

$$e^{2\overline{a}(1-t)^{3/2}/3}(1-t)^2 E_{-1/3}\left(\frac{2\overline{a}(1-t)^{3/2}}{3}\right) \leq \frac{3}{\overline{a}}\sqrt{1-t} \leq \frac{3}{\overline{a}}, \\ e^{2\overline{a}(1-t)^{3/2}/3}(1-t) E_{1/3}\left(\frac{2\overline{a}(1-t)^{3/2}}{3}\right) \leq \frac{3}{\overline{a}}(1-t)^{-1/2} \leq \frac{3}{\overline{a}^{2/3}}. \tag{J.155}$$

Instead, if $2\overline{a}(1-t)^{3/2}/3 < 2/3$, by using (J.110) we have

$$e^{2\overline{a}(1-t)^{3/2}/3}(1-t)^2 E_{-1/3}\left(\frac{2\overline{a}(1-t)^{3/2}}{3}\right) \leq e^{2/3}\Gamma\left(\frac{4}{3}\right)\left(\frac{2\overline{a}}{3}\right)^{-4/3}, \\ e^{2\overline{a}(1-t)^{3/2}/3}(1-t) E_{1/3}\left(\frac{2\overline{a}(1-t)^{3/2}}{3}\right) \leq e^{2/3}\Gamma\left(\frac{2}{3}\right)\left(\frac{2\overline{a}}{3}\right)^{-2/3}. \tag{J.156}$$

Thus, the following upper bound holds for $t \in [0,1]$:

$$\overline{R}'(t) \leq R(0) + \frac{2\overline{b}_1}{\overline{a}} + \frac{2\overline{b}_2}{\overline{a}^{2/3}} + e^{2/3}\left(\frac{2}{3}\right)^{-1/3}\Gamma\left(\frac{4}{3}\right)\frac{\overline{b}_1}{\overline{a}^{4/3}} + e^{2/3}\left(\frac{2}{3}\right)^{1/3}\Gamma\left(\frac{2}{3}\right)\frac{\overline{b}_2}{\overline{a}^{2/3}}. \tag{J.157}$$

Following the same passages, we have that, for $t \in [0,1]$,

$$\underline{R}'(t) \leq R(0) + \frac{2\underline{b}_1}{\underline{a}} + \frac{2\underline{b}_2}{\underline{a}^{2/3}} + e^{2/3}\left(\frac{2}{3}\right)^{-1/3}\Gamma\left(\frac{4}{3}\right)\frac{\underline{b}_1}{\underline{a}^{4/3}} + e^{2/3}\left(\frac{2}{3}\right)^{1/3}\Gamma\left(\frac{2}{3}\right)\frac{\underline{b}_2}{\underline{a}^{2/3}}. \tag{J.158}$$

Let us now pick

$$\overline{a} = 2vc\lambda_{\min}\frac{\underline{c}_\mu(c,\zeta)}{\sqrt{R(0)+1+\zeta^2/2}}, \qquad \overline{b}_1 = \lambda_{\max}\frac{v^2c^2\gamma}{2}, \qquad \overline{b}_2 = 4v^2c^2\frac{\gamma^2}{\rho^2}, \\ \underline{a} = 2vc\lambda_{\max}\frac{1}{\sqrt{\pi\zeta^2/2}}, \qquad \underline{b}_1 = \frac{v^2c^2\gamma\underline{c}_\nu(c,\zeta)}{2}, \qquad \underline{b}_2 = v^2c^2\frac{\gamma^2}{\rho^2}. \tag{J.159}$$

Since $c = O(1)$, by Lemma J.5, we have that (J.128) holds. As $cv = C\ln(1/\gamma)/\lambda_{\min}$, the choice in (J.159) ensures that $0 \le \underline{R}'(t) \le \overline{R}'(t) \le 1 + R(0)$ for $t \in [0,1]$, as long as we have $\overline{b}_1 + \overline{b}_2 = o(1)$, which holds due to (J.146).

Thus, (J.159) guarantees

$$\overline{f}(t, \overline{R}'(t)) < \overline{f}'(t, \overline{R}'(t)), \qquad \underline{f}(t, \underline{R}'(t)) > \underline{f}'(t, \underline{R}'(t)), \qquad \text{for all } \ t \in [0,1]. \tag{J.160}$$

Note that $\overline{f}(t,R)$ and $\underline{f}(t,R)$ are continuous in both variables in the intervals $t \in [0,1]$ and $R \in [0,1]$. Furthermore, they are also Lipschitz in R in these same intervals. Thus, (J.106) gives

$$\underline{R}'(t) < \underline{R}(t), \qquad \overline{R}(t) < \overline{R}'(t), \qquad \text{for all } \ t \in (0,1]. \tag{J.161}$$

To prove the upper bound, due to Theorem 9.6 and Propositions 9.8 and J.12, it suffices to give an upper bound on $\overline{R}'(1)$. To this aim, note that Lemma J.13 yields

$$\begin{aligned} \lim_{x\to 0} x^2\, E_{-1/3}\left(\frac{2ax^{3/2}}{3}\right) &= \left(\frac{3}{2}\right)^{4/3} \Gamma\left(\frac{4}{3}\right) a^{-4/3}, \\ \lim_{x\to 0} x\, E_{1/3}\left(\frac{2ax^{3/2}}{3}\right) &= \left(\frac{3}{2}\right)^{2/3} \Gamma\left(\frac{2}{3}\right) a^{-2/3}, \end{aligned} \tag{J.162}$$

where $\Gamma(\cdot)$ denotes the Euler Gamma function (defined in (J.109)). Thus,

$$\begin{aligned} \overline{R}'(1) &= \frac{R(0)}{3} e^{-2\overline{a}/3} \left(3 - \frac{2}{R(0)} \overline{b}_1 e^{2\overline{a}/3} \left(E_{-1/3}\left(\frac{2\overline{a}}{3}\right) - \left(\frac{3}{2}\right)^{4/3} \Gamma\left(\frac{4}{3}\right) \overline{a}^{-4/3} \right) \right. \\ &\qquad \left. - \frac{2}{R(0)} \overline{b}_2 e^{2\overline{a}/3} \left(E_{1/3}\left(\frac{2\overline{a}}{3}\right) - \left(\frac{3}{2}\right)^{2/3} \Gamma\left(\frac{2}{3}\right) \overline{a}^{-2/3} \right) \right) \\ &\le R(0) e^{-2\overline{a}/3} + \left(\frac{3}{2}\right)^{1/3} \Gamma\left(\frac{4}{3}\right) \overline{b}_1 \overline{a}^{-4/3} + \left(\frac{3}{2}\right)^{-1/3} \Gamma\left(\frac{2}{3}\right) \overline{b}_2 \overline{a}^{-2/3}, \end{aligned} \tag{J.163}$$

where in the last line we have used the non-negativity of the exponential integral functions. For C sufficiently large, due to (J.159), the first term in the RHS is $o(\gamma)$ (recall that $vc = C\ln(1/\gamma)/\lambda_{\min}$). The other two terms are $O(\lambda_{\max}\gamma \ln^{2/3}(1/\gamma)/\lambda_{\min}^2)$ and $O(\gamma^2 \ln^{4/3}(1/\gamma)/(\rho^2\lambda_{\min}^2)$ respectively. Note that $\overline{R}(1) < \overline{R}'(1)$ and $\tilde{\eta}(1) = 0$. Thus, the desired result follows from Theorem 9.6 and Propositions 9.8 and J.12.

To prove the lower bound, due to Theorem 9.6 and Propositions 9.8 and J.12, it suffices to show that the inequality in the thesis holds for $\underline{R}(1)$. Let us first suppose $c \le 1$, which implies that (J.128) holds. Pick $\underline{a}, \underline{b}_1, \underline{b}_2$ as in (J.159), and consider the ODE $\underline{R}'(t)$ defined in (J.151) with the initial condition $\underline{R}'(0) = R(0)$, which is a lower bound on $\underline{R}(t)$ due to (J.161), i.e.,

$$\begin{aligned} \underline{R}(1) > \underline{R}'(1) = \frac{R(0)}{3} e^{-2\underline{a}/3} \left(3 - \frac{2}{R(0)} \underline{b}_1 e^{2\underline{a}/3} \left(E_{-1/3}\left(\frac{2\underline{a}}{3}\right) - \left(\frac{3}{2}\right)^{4/3} \Gamma\left(\frac{4}{3}\right) \underline{a}^{-4/3} \right) \right. \\ \left. - \frac{2}{R(0)} \underline{b}_2 e^{2\underline{a}/3} \left(E_{1/3}\left(\frac{2\underline{a}}{3}\right) - \left(\frac{3}{2}\right)^{2/3} \Gamma\left(\frac{2}{3}\right) \underline{a}^{-2/3} \right) \right). \end{aligned} \tag{J.164}$$

We consider two additional cases depending on the value of $\underline{a}$. If $\underline{a} \ge 2$, then (J.111) gives that

$$E_{-1/3}\left(\frac{2\underline{a}}{3}\right) \le \frac{2e^{-2\underline{a}/3}}{\frac{2\underline{a}}{3}}, \qquad E_{1/3}\left(\frac{2\underline{a}}{3}\right) \le \frac{2e^{-2\underline{a}/3}}{\frac{2\underline{a}}{3}}, \tag{J.165}$$

which implies that

$$\begin{aligned}\underline{R}(1) \geq R(0)e^{-2\underline{a}/3} + \left(\frac{3}{2}\right)^{1/3}\Gamma\left(\frac{4}{3}\right)\underline{b}_1\underline{a}^{-4/3} - 2e^{-2\underline{a}/3}\underline{b}_1\underline{a}^{-1} \\ + \left(\frac{3}{2}\right)^{-1/3}\Gamma\left(\frac{2}{3}\right)\underline{b}_2\underline{a}^{-2/3} - 2e^{-2\underline{a}/3}\underline{b}_2\underline{a}^{-1}.\end{aligned} \tag{J.166}$$

Note that

$$\begin{aligned}\left(\frac{3}{2}\right)^{1/3}\Gamma\left(\frac{4}{3}\right)\underline{b}_1\underline{a}^{-4/3} - 2e^{-2\underline{a}/3}\underline{b}_1\underline{a}^{-1} \geq \underline{b}_1\underline{a}^{-4/3} - 2e^{-2\underline{a}/3}\underline{b}_1\underline{a}^{-1} \geq \frac{1}{3}\underline{b}_1\underline{a}^{-4/3}, \\ \left(\frac{3}{2}\right)^{-1/3}\Gamma\left(\frac{2}{3}\right)\underline{b}_2\underline{a}^{-2/3} - 2e^{-2\underline{a}/3}\underline{b}_2\underline{a}^{-1} \geq \underline{b}_2\underline{a}^{-2/3} - 2e^{-2\underline{a}/3}\underline{b}_2\underline{a}^{-1} \geq \frac{1}{3}\underline{b}_2\underline{a}^{-2/3},\end{aligned} \tag{J.167}$$

where the inequalities on the right hold for $\underline{a} \geq 2$. Thus,

$$\begin{aligned}\underline{R}(1) \geq \left(\frac{R(0)}{2}e^{-avc\lambda_{\max}} + \frac{b_1}{3}\lambda_{\max}^{-4/3}a^{-4/3}(vc)^{2/3}\gamma\right) \\ + \left(\frac{R(0)}{2}e^{-avc\lambda_{\max}} + \frac{b_2}{3}\lambda_{\max}^{-2/3}a^{-2/3}(vc)^{4/3}\frac{\gamma^2}{\rho^2}\right),\end{aligned} \tag{J.168}$$

where a, b_1, b_2 are positive constants which do not depend on γ, v, c, ρ or the spectrum of Σ. Denoting with $a' = a\lambda_{\max}$, $\tau = vc$, we have that for a fixed $\beta \in \{2/3, 4/3\}$, and for any $\delta = o(1)$,

$$\min_{\tau \geq 0} e^{-\tau a'} + \frac{\tau^{2-\beta}\delta}{(a')^\beta} = \Omega\left(\frac{\delta \ln^{2-\beta}(1/\delta)}{(a')^2}\right), \tag{J.169}$$

which follows from the following calculations:

$$\begin{aligned}\min_{\tau \geq \ln(1/\delta)/(2a')} e^{-\tau a'} + \frac{\tau^{2-\beta}\delta}{(a')^\beta} \geq \min_{\tau \geq \ln(1/\delta)/(2a')} \frac{\tau^{2-\beta}\delta}{(a')^\beta} = \Omega\left(\frac{\delta \ln^{2-\beta}(1/\delta)}{(a')^2}\right), \\ \min_{0 \leq \tau \leq \ln(1/\delta)/(2a')} e^{-\tau a'} + \frac{\tau^{2-\beta}\delta}{(a')^\beta} \geq \delta^{1/2} = \Omega(\delta \ln^{2-\beta}(1/\delta)) = \Omega\left(\frac{\delta \ln^{2-\beta}(1/\delta)}{(a')^2}\right).\end{aligned} \tag{J.170}$$

Then, since $\rho = \Omega(\gamma^{1-h})$, we can set $\delta = \gamma$ and $\delta = \gamma^2/\rho^2$ to obtain

$$\underline{R}(1) = \Omega\left(\frac{\gamma \ln^{2/3}(1/\gamma)}{\lambda_{\max}^2} + \frac{\gamma^2 \ln^{4/3}(\rho^2/\gamma^2)}{\rho^2\lambda_{\max}^2}\right), \tag{J.171}$$

which implies the desired result (due to Theorem 9.6 and Propositions 9.8 and J.12).

If $\underline{a} \leq 2$, then (J.110) gives that

$$E_{-1/3}\left(\frac{2\underline{a}}{3}\right) \leq \Gamma\left(\frac{4}{3}\right)\left(\frac{2\underline{a}}{3}\right)^{-4/3}, \qquad E_{1/3}\left(\frac{2\underline{a}}{3}\right) \leq \Gamma\left(\frac{2}{3}\right)\left(\frac{2\underline{a}}{3}\right)^{-2/3}, \tag{J.172}$$

which implies that

$$\underline{R}(1) \geq R(0)e^{-2\underline{a}/3} = \Omega(1), \tag{J.173}$$

thus again proving the desired claim (due to Theorem 9.6 and Propositions 9.8 and J.12).

Finally, for $c > 1$, due to the same argument used to show (J.142), the solution of the original ODE is lower bounded by that of the ODE below:

$$\mathsf{d}\underline{R}'' = -2\lambda_{\max} v\sqrt{1-t}\underline{R}''\mathsf{d}t + v^2 b_1 \gamma(1-t)\mathsf{d}t + 2v^2c^2\frac{\gamma^2}{\rho^2}\mathsf{d}t, \tag{J.174}$$

with initial condition $\underline{R}''(0) = R(0)$, where b_1 is a positive constant independent of γ, v, c, ρ and the spectrum of Σ. Thus, following the same steps above with $\underline{a} = 2v\lambda_{\max}, \underline{b}_1 = v^2 b_1 \gamma, \underline{b}_2 = 2v^2c^2\gamma^2/\rho^2$, one readily shows that

$$\underline{R}''(1) = \Omega\left(\frac{\gamma \ln^{2/3}(1/\gamma)}{\lambda_{\max}^2} + \frac{\gamma^2 \ln^{4/3}(1/\gamma)}{\rho^2\lambda_{\max}^2}\right), \tag{J.175}$$

concluding the proof (due to Theorem 9.6 and Propositions 9.8 and J.12). □

Sub-optimality of $c = \omega_\gamma(1)$. Note that, if $c > 1$, the ODE in (J.143) maps to the one in (J.127), with $\underline{a}, \underline{b}$ defined as in (J.129) (except for absolute constants), with $vc \mapsto v$. This mapping also holds when defining (J.136), via $\rho \mapsto \rho/c$. Then, if we consider the further condition $\rho/c = \Omega(\gamma^{1-h})$, for some $h > 0$, the lower bound in Proposition 9.9 becomes

$$\mathcal{R}(\theta_0^p) = \Omega\left(\frac{\gamma \ln(1/\gamma)}{\lambda_{\max}^2} + \frac{\max(1, c^2)\gamma^2 \ln^2(1/\gamma)}{\rho^2\lambda_{\max}^2}\right). \tag{J.176}$$

The same argument holds also for the ODE in (J.174), providing analogous expression for $\mathcal{R}(\theta_{1/2}^p)$. This quantifies the negative effects of using a large clipping constant $c = \omega_\gamma(1)$.

J.4.2 Proofs on decaying private noise: $\alpha \geq 1$

In this section, we prove Proposition 9.10. The result is stated in the general case where $\lambda_{\max}$ and $\lambda_{\min}$ are allowed to depend on γ.

Theorem. General version of Proposition 9.10. Let Assumptions 9.3 and 9.4 hold, and let θ_α^p be the solution obtained with Algorithm 9.1, with $\tilde{\eta}(t)$ given by (9.33) for $\alpha \geq 1$. Consider the setting

$$\gamma = \frac{d}{n} = o_\gamma(1), \qquad \frac{\ln^2(1/\gamma)\gamma}{\lambda_{\min}^2}\left(\lambda_{\max}\alpha + \frac{\gamma}{\rho^2}\right) = o_\gamma(1), \tag{J.177}$$

and pick

$$c = O_\gamma(1), \qquad \tilde{\eta}(0)c = \frac{C\alpha\ln(1/\gamma)}{\lambda_{\min}}, \qquad \tilde{\eta}(0) \leq \frac{2}{\gamma}, \tag{J.178}$$

for a large enough constant C which does not depend on $\gamma, \rho, \Sigma, \alpha$. Then, we have that, with overwhelming probability,

$$\mathcal{R}(\theta_\alpha^p) = O_\gamma\left(\frac{\lambda_{\max}}{\lambda_{\min}^2}\alpha\gamma\ln^{\frac{1}{1+\alpha}}(1/\gamma) + \frac{1}{\lambda_{\min}^2}\frac{\alpha^2\gamma^2\ln^{\frac{2}{1+\alpha}}(1/\gamma)}{\rho^2}\right). \tag{J.179}$$

Proof. Let $\overline{R}(t)$ be defined as in (J.99), with the learning rate schedule in (9.33) for a generic $\alpha \geq 1$. Then, we have

$$\begin{aligned}\mathsf{d}\overline{R} = &-2vc\lambda_{\min}(1-t)^\alpha\frac{\mu_c(\overline{R})}{c}\overline{R}\mathsf{d}t + (vc)^2\lambda_{\max}(1-t)^{2\alpha}\frac{\nu_c(\overline{R})(\overline{R}+\zeta^2/2)}{c^2}\gamma\mathsf{d}t \\ &+ 4(vc)^2\alpha(1-t)^{2\alpha-1}\frac{\gamma^2}{\rho^2}\mathsf{d}t.\end{aligned} \tag{J.180}$$

The argument is similar to the one used to obtain Proposition 9.9, so we only highlight differences. Using (J.106), we have that $\overline{R}(t)$ is strictly bounded by the auxiliary ODEs

$$\begin{aligned}
\mathsf{d}\overline{R}' &= -\overline{a}(1-t)^{\alpha}\overline{R}'\mathsf{d}t + \overline{b}_1(1-t)^{2\alpha}\mathsf{d}t + \overline{b}_2(1-t)^{2\alpha-1}\mathsf{d}t =: \overline{f}'(t,\overline{R}')\mathsf{d}t, \qquad \overline{a},\overline{b}_1,\overline{b}_2 > 0,\\
\mathsf{d}\underline{R}' &= -\underline{a}(1-t)^{\alpha}\underline{R}'\mathsf{d}t + \underline{b}_1(1-t)^{2\alpha}\mathsf{d}t + \underline{b}_2(1-t)^{2\alpha-1}\mathsf{d}t =: \underline{f}'(t,\underline{R}')\mathsf{d}t, \qquad \underline{a},\underline{b}_1,\underline{b}_2 > 0,
\end{aligned} \tag{J.181}$$

with initial conditions $\overline{R}'(0) = \overline{R}(0) = R(0)$.

To establish a closed form solution for the ODEs in (J.181), we start by analyzing the ODE

$$\mathsf{d}\tilde{R} = -\overline{a}(1-t)^{\alpha}\tilde{R}\mathsf{d}t + \overline{b}_1(1-t)^{2\alpha}\mathsf{d}t, \tag{J.182}$$

with initial condition $\tilde{R}(0) = R(0)$, which admits the closed form solution

$$\begin{aligned}
\tilde{R}(t) = \frac{R(0)}{(1+\alpha)}e^{-\frac{\overline{a}(1-(1-t)^{1+\alpha})}{1+\alpha}}\Bigg(1+\alpha - \frac{\overline{b}_1}{R(0)}e^{\frac{\overline{a}}{1+\alpha}}E_{-1+\frac{1}{1+\alpha}}\left(\frac{\overline{a}}{1+\alpha}\right)\\
+ \frac{\overline{b}_1}{R(0)}e^{\frac{\overline{a}}{1+\alpha}}(1-t)^{1+2\alpha}E_{-1+\frac{1}{1+\alpha}}\left(\frac{\overline{a}(1-t)^{1+\alpha}}{1+\alpha}\right)\Bigg).
\end{aligned} \tag{J.183}$$

Let

$$w(t) = \overline{R}'(t) - \tilde{R}(t), \tag{J.184}$$

and note that

$$\begin{aligned}
\frac{\mathsf{d}w}{\mathsf{d}t} &= \frac{\mathsf{d}\overline{R}'}{\mathsf{d}t} - \frac{\mathsf{d}\tilde{R}}{\mathsf{d}t}\\
&= -\overline{a}(1-t)^{\alpha}\overline{R} + \overline{a}(1-t)^{\alpha}\tilde{R} + \overline{b}_2(1-t)^{2\alpha-1}\\
&= -\overline{a}(1-t)^{\alpha}w + \overline{b}_2(1-t)^{2\alpha-1}.
\end{aligned} \tag{J.185}$$

Thus, by using the initial condition $w(0) = \overline{R}'(0) - \tilde{R}(0) = 0$, we have

$$w(t) = \frac{\overline{b}_2}{1+\alpha}e^{\frac{\overline{a}(1-t)^{1+\alpha}}{1+\alpha}}\left((1-t)^{2\alpha}E_{\frac{1-\alpha}{1+\alpha}}\left(\frac{\overline{a}(1-t)^{1+\alpha}}{1+\alpha}\right) - E_{\frac{1-\alpha}{1+\alpha}}\left(\frac{\overline{a}}{1+\alpha}\right)\right), \tag{J.186}$$

which implies that

$$\begin{aligned}
\overline{R}'(t) = \frac{R(0)}{(1+\alpha)}e^{-\frac{\overline{a}(1-(1-t)^{1+\alpha})}{1+\alpha}}\Bigg(1+\alpha - \frac{\overline{b}_1}{R(0)}e^{\frac{\overline{a}}{1+\alpha}}E_{-1+\frac{1}{1+\alpha}}\left(\frac{\overline{a}}{1+\alpha}\right)\\
+ \frac{\overline{b}_1}{R(0)}e^{\frac{\overline{a}}{1+\alpha}}(1-t)^{1+2\alpha}E_{-1+\frac{1}{1+\alpha}}\left(\frac{\overline{a}(1-t)^{1+\alpha}}{1+\alpha}\right)\\
+\frac{\overline{b}_2}{R(0)}e^{\frac{\overline{a}}{1+\alpha}}\left((1-t)^{2\alpha}E_{\frac{1-\alpha}{1+\alpha}}\left(\frac{\overline{a}(1-t)^{1+\alpha}}{1+\alpha}\right) - E_{\frac{1-\alpha}{1+\alpha}}\left(\frac{\overline{a}}{1+\alpha}\right)\right)\Bigg).
\end{aligned} \tag{J.187}$$

Similarly, we have that

$$\begin{aligned}
\underline{R}'(t) = \frac{1}{(1+\alpha)}e^{-\frac{\underline{a}(1-(1-t)^{1+\alpha})}{1+\alpha}}\Bigg(R(0)(1+\alpha) - \underline{b}_1e^{\frac{\underline{a}}{1+\alpha}}E_{-1+\frac{1}{1+\alpha}}\left(\frac{\underline{a}}{1+\alpha}\right)\\
+ \underline{b}_1e^{\frac{\underline{a}}{1+\alpha}}(1-t)^{1+2\alpha}E_{-1+\frac{1}{1+\alpha}}\left(\frac{\underline{a}(1-t)^{1+\alpha}}{1+\alpha}\right)\\
+\underline{b}_2e^{\frac{\underline{a}}{1+\alpha}}\left((1-t)^{2\alpha}E_{\frac{1-\alpha}{1+\alpha}}\left(\frac{\underline{a}(1-t)^{1+\alpha}}{1+\alpha}\right) - E_{\frac{1-\alpha}{1+\alpha}}\left(\frac{\underline{a}}{1+\alpha}\right)\right)\Bigg).
\end{aligned} \tag{J.188}$$

For $t \in [0,1]$, $\overline{R}(t), \underline{R}(t) \geq 0$. Furthermore, the following upper bounds hold for $t \in [0,1]$:

$$\begin{aligned}\overline{R}'(t) \leq R(0) + \frac{2\bar{b}_1}{\bar{a}} + \frac{\bar{b}_2}{\alpha} + \bar{b}_1 \left(\frac{1}{1+\alpha}\right)^{-1+\frac{1}{1+\alpha}} \Gamma\left(2 - \frac{1}{1+\alpha}\right) \bar{a}^{-2+\frac{1}{1+\alpha}} \\ + \bar{b}_2 \Gamma\left(\frac{2\alpha}{1+\alpha}\right) (1+\alpha)^{\frac{\alpha-1}{1+\alpha}} \bar{a}^{-\frac{2\alpha}{1+\alpha}}.\end{aligned} \tag{J.189}$$

Let us take

$$\bar{a} = C_1 v c \lambda_{\min}, \qquad \bar{b}_1 = C_2 \gamma v^2 c^2 \lambda_{\max}, \qquad \bar{b}_2 = C_3 v^2 c^2 \frac{\gamma^2}{\rho^2} \alpha, \tag{J.190}$$

with C_1, C_2, C_3 constants independent from $\gamma, \rho, \alpha, v, c$ or the spectrum of Σ. Then, we have that $\overline{R}'(t) \leq 1 + R(0)$ for $t \in [0,1]$, as $\bar{a} = \Omega(\ln(1/\gamma))$ and

$$\frac{\ln^2(1/\gamma)\gamma}{\lambda_{\min}^2} \left(\lambda_{\max}\alpha + \frac{\gamma}{\rho^2}\right) = o(1). \tag{J.191}$$

As a result, we can apply the argument in (J.106) to obtain that

$$\overline{R}(1) \leq \overline{R}'(1) = \tilde{R}(1) + w(1). \tag{J.192}$$

Note that Lemma J.13 yields

$$\lim_{x \to 0} x^{1+2\alpha} E_{-1+\frac{1}{1+\alpha}}\left(\frac{\bar{a}x^{1+\alpha}}{1+\alpha}\right) = \left(\frac{1}{1+\alpha}\right)^{-2+\frac{1}{1+\alpha}} \Gamma\left(2 - \frac{1}{1+\alpha}\right) \bar{a}^{-2+\frac{1}{1+\alpha}}. \tag{J.193}$$

Thus,

$$\begin{aligned}\tilde{R}(1) &= \frac{1}{(1+\alpha)} e^{-\frac{\bar{a}}{1+\alpha}} \Bigg(R(0)(1+\alpha) - \bar{b}_1 e^{\frac{\bar{a}}{1+\alpha}} E_{-1+\frac{1}{1+\alpha}}\left(\frac{\bar{a}}{1+\alpha}\right) \\ &\qquad + \bar{b}_1 e^{\frac{\bar{a}}{1+\alpha}} \left(\frac{1}{1+\alpha}\right)^{-2+\frac{1}{1+\alpha}} \Gamma\left(2 - \frac{1}{1+\alpha}\right) \bar{a}^{-2+\frac{1}{1+\alpha}} \Bigg) \\ &\leq R(0) e^{-\frac{\bar{a}}{1+\alpha}} + \left(\frac{1}{1+\alpha}\right)^{-1+\frac{1}{1+\alpha}} \Gamma\left(2 - \frac{1}{1+\alpha}\right) \bar{b}_1 \bar{a}^{-2+\frac{1}{1+\alpha}} \\ &\leq R(0) e^{-\frac{\bar{a}}{1+\alpha}} + (1+\alpha) \bar{b}_1 \bar{a}^{-2+\frac{1}{1+\alpha}},\end{aligned} \tag{J.194}$$

where in the second line we have used the non-negativity of the exponential integral function and in the third line we have used that $\Gamma\left(2 - \frac{1}{1+\alpha}\right) \leq \Gamma(2) = 1$ for all $\alpha \geq 1$. Next, we bound $w(1)$ as

$$\begin{aligned}w(1) &= \frac{\bar{b}_2}{1+\alpha} \left(\Gamma\left(\frac{2\alpha}{1+\alpha}\right) (1+\alpha)^{\frac{2\alpha}{1+\alpha}} \bar{a}^{-\frac{2\alpha}{1+\alpha}} - E_{\frac{1-\alpha}{1+\alpha}}\left(\frac{\bar{a}}{1+\alpha}\right)\right) \\ &\leq \frac{\bar{b}_2}{1+\alpha} \Gamma\left(\frac{2\alpha}{1+\alpha}\right) (1+\alpha)^{\frac{2\alpha}{1+\alpha}} \bar{a}^{-\frac{2\alpha}{1+\alpha}} \\ &\leq (1+\alpha) \bar{b}_2 \bar{a}^{-\frac{2\alpha}{1+\alpha}},\end{aligned} \tag{J.195}$$

where in the last line we have used that $\Gamma\left(\frac{2\alpha}{1+\alpha}\right) \leq 1$ for $\alpha \geq 1$. By combining (J.192), (J.194) and (J.195), we conclude that

$$\begin{aligned}R(1) &\leq R(0) e^{-\frac{\bar{a}}{1+\alpha}} + (1+\alpha)^{1/3} \bar{b}_1 \bar{a}^{-\frac{2\alpha+1}{1+\alpha}} + (1+\alpha) \bar{b}_2 \bar{a}^{-\frac{2\alpha}{1+\alpha}} \\ &= O\left(\frac{\lambda_{\max}\alpha\gamma \ln^{\frac{1}{1+\alpha}}(1/\gamma)}{\lambda_{\min}^2} + \frac{\alpha^2\gamma^2 \ln^{\frac{2}{1+\alpha}}(1/\gamma)}{\rho^2 \lambda_{\min}^2}\right),\end{aligned} \tag{J.196}$$

which gives the thesis, after applying Theorem 9.6 and Propositions 9.8 and J.12, since $\tilde{\eta}(1) = 0$. □

Proof of (9.49)**.** By taking $\alpha = \ln\ln(1/\gamma)$, we have

$$\begin{aligned}
&\alpha\gamma \ln^{\frac{1}{1+\alpha}}(1/\gamma) + \frac{\alpha^2\gamma^2 \ln^{\frac{2}{1+\alpha}}(1/\gamma)}{\rho^2} \\
&\quad = \gamma \ln\ln(1/\gamma) e^{\frac{\ln\ln(1/\gamma)}{1+\ln\ln(1/\gamma)}} + \frac{\gamma^2}{\rho^2}(\ln\ln(1/\gamma))^2 e^{\frac{2\ln\ln(1/\gamma)}{1+\ln\ln(1/\gamma)}} \qquad \text{(J.197)} \\
&\quad \le \gamma(\ln\ln(1/\gamma))e + \frac{\gamma^2}{\rho^2}(\ln\ln(1/\gamma))^2 e^2,
\end{aligned}$$

which readily gives (9.49), as $\lambda_{\max}, \lambda_{\min} = \Theta_\gamma(1)$. □

J.4.3 Proofs on the harmonic schedule

In this section, we prove Theorem 9.11, providing a more general result that allows for $\lambda_{\max}$ and $\lambda_{\min}$ to depend on γ.

Lemma J.16. *Consider the ODE*

$$\mathsf{d}\tilde{R} = -a\tilde{\eta}(t)\tilde{R}\mathsf{d}t + b_1\tilde{\eta}^2(t)\mathsf{d}t - b_2\frac{\mathsf{d}\tilde{\eta}^2(t)}{\mathsf{d}t}\mathsf{d}t, \qquad a, b_1, b_2 > 0, \qquad \text{(J.198)}$$

with initial condition $\tilde{R}(0)$. Then,

$$\tilde{R}(t) = e^{-aF(t)}\left(\tilde{R}(0) + \int_0^t e^{aF(s)}\Big(b_1\tilde{\eta}^2(s) - 2b_2\tilde{\eta}(s)\tilde{\eta}'(s)\Big)\mathsf{d}s\right), \qquad \text{(J.199)}$$

where $\tilde{\eta}'(t)$ denotes the derivative of $\tilde{\eta}(t)$ and $F(t) := \int_0^t \tilde{\eta}(s)\mathsf{d}s$.

Proof. Define

$$u(t) := \tilde{R}(t) + b_2\tilde{\eta}^2(t).$$

Then,

$$\mathsf{d}u(t) = -a\tilde{\eta}(t)u(t)\mathsf{d}t + b_1\tilde{\eta}^2(t)\mathsf{d}t + ab_2\tilde{\eta}^3(t)\mathsf{d}t,$$

which implies

$$\mathsf{d}\left(u(t)e^{aF(t)}\right) = b_1\tilde{\eta}^2(t)e^{aF(t)}\mathsf{d}t + ab_2\tilde{\eta}^3(t)e^{aF(t)}\mathsf{d}t.$$

By integrating the equation above, we obtain

$$u(t)e^{aF(t)} = u(0)e^{aF(0)} + \int_0^t e^{aF(s)}\Big(b_1\tilde{\eta}^2(s) + ab_2\tilde{\eta}^3(s)\Big)\mathsf{d}s.$$

Noting that $u(0) = \tilde{R}(0) + b_2\tilde{\eta}^2(0)$ and $F(0) = 0$ gives

$$\tilde{R}(t) = -b_2\tilde{\eta}^2(t) + e^{-aF(t)}\tilde{R}(0) + e^{-aF(t)}b_2\tilde{\eta}^2(0) + e^{-aF(t)}\int_0^t e^{aF(s)}\Big(b_1\tilde{\eta}^2(s) + ab_2\tilde{\eta}^3(s)\Big)\mathsf{d}s. \qquad \text{(J.200)}$$

Furthermore, we have

$$\frac{\mathsf{d}}{\mathsf{d}s}\left(e^{aF(s)}\tilde{\eta}^2(s)\right) = ae^{aF(s)}\tilde{\eta}^3(s) + 2e^{aF(s)}\tilde{\eta}(s)\tilde{\eta}'(s),$$

which implies that

$$\int_0^t ae^{aF(s)}\tilde{\eta}^3(s)\mathsf{d}s = e^{aF(t)}\tilde{\eta}^2(t) - \tilde{\eta}^2(0) - 2\int_0^t e^{aF(s)}\tilde{\eta}(s)\tilde{\eta}'(s)\mathsf{d}s. \qquad \text{(J.201)}$$

Combining (J.200) and (J.201) gives the desired result. □

Theorem. General version of Theorem 9.11. Let Assumptions 9.3 and 9.4 hold, and let θ^p be the solution obtained with Algorithm 9.1. Assume that

$$\frac{\lambda_{\max}}{\lambda_{\min}^2}\gamma + \frac{1}{\lambda_{\min}^2}\frac{\gamma^2}{\rho^2} = o_\gamma(1). \tag{J.202}$$

Consider the schedule

$$\tilde{\eta}(t) = \frac{\alpha}{t+\tau}, \qquad \text{with } \alpha := \frac{3\sqrt{A+90}}{\lambda_{\min}}, \quad \tau := \max\left(\frac{9(A+90)\lambda_{\max}}{\lambda_{\min}^2}\gamma, \frac{12\sqrt{2(A+90)}}{\lambda_{\min}}\frac{\gamma}{\rho}\right), \tag{J.203}$$

where A is a universal constant s.t. $R(0)+\zeta^2/2 \leq A$. Then, by setting $c = 3\sqrt{2(A+90)}/\lambda_{\min}$, we have that, with overwhelming probability,

$$\mathcal{R}(\theta^p) = O_\gamma\left(\frac{\lambda_{\max}}{\lambda_{\min}^4}\gamma + \frac{1}{\lambda_{\min}^4}\frac{\gamma^2}{\rho^2}\right). \tag{J.204}$$

Proof. Let us introduce the shorthands

$$\begin{aligned}
\overline{f}(t,\overline{R}) &= -2\lambda_{\min}\tilde{\eta}(t)\mu_c(\overline{R})\overline{R} + \lambda_{\max}c^2\tilde{\eta}^2(t)\frac{\nu_c(\overline{R})(\overline{R}+\zeta^2/2)}{c^2}\gamma - 2c^2\frac{\gamma^2}{\rho^2}\frac{\mathsf{d}\tilde{\eta}^2(t)}{\mathsf{d}t},\\
\underline{f}(t,\underline{R}) &= -2\lambda_{\max}\tilde{\eta}(t)\mu_c(\underline{R})\underline{R} + c^2\tilde{\eta}^2(t)\frac{\nu_c(\underline{R})(\underline{R}+\zeta^2/2)}{c^2}\gamma - 2c^2\frac{\gamma^2}{\rho^2}\frac{\mathsf{d}\tilde{\eta}^2(t)}{\mathsf{d}t}.
\end{aligned}$$

Then, consider the auxiliary ODEs

$$\begin{aligned}
\mathsf{d}\overline{R}' &= -\overline{a}\tilde{\eta}(t)\overline{R}'\mathsf{d}t + \overline{b}_1\tilde{\eta}^2(t)\mathsf{d}t - \overline{b}_2\frac{\mathsf{d}\tilde{\eta}^2(t)}{\mathsf{d}t}\mathsf{d}t =: \overline{f}'(t,\overline{R}')\mathsf{d}t, \qquad \overline{a},\overline{b}_1,\overline{b}_2 > 0,\\
\mathsf{d}\underline{R}' &= -\underline{a}\tilde{\eta}(t)\underline{R}'\mathsf{d}t + \underline{b}_1\tilde{\eta}^2(t)\mathsf{d}t - \underline{b}_2\frac{\mathsf{d}\tilde{\eta}^2(t)}{\mathsf{d}t}\mathsf{d}t =: \underline{f}'(t,\underline{R}')\mathsf{d}t, \qquad \underline{a},\underline{b}_1,\underline{b}_2 > 0,
\end{aligned} \tag{J.205}$$

with initial conditions $\overline{R}'(0) = \overline{R}(0) = R(0) = \underline{R}(0) = \underline{R}'(0)$. By Lemma J.16, we have

$$\begin{aligned}
\overline{R}'(t) &= e^{-\overline{a}F(t)}\left(R(0) + \int_0^t e^{\overline{a}F(s)}\left(\overline{b}_1\tilde{\eta}^2(s) - 2\overline{b}_2\tilde{\eta}(s)\tilde{\eta}'(s)\right)\mathsf{d}s\right),\\
\underline{R}'(t) &= e^{-\underline{a}F(t)}\left(R(0) + \int_0^t e^{\underline{a}F(s)}\left(\underline{b}_1\tilde{\eta}^2(s) - 2\underline{b}_2\tilde{\eta}(s)\tilde{\eta}'(s)\right)\mathsf{d}s\right),
\end{aligned}$$

where

$$F(t) := \int_0^t \tilde{\eta}(s)\mathsf{d}s.$$

By plugging in the schedule defined in (J.203), we obtain

$$F(t) = \alpha\log\frac{t+\tau}{\tau}, \qquad \tilde{\eta}'(t) = -\frac{\alpha}{(t+\tau)^2}.$$

Thus, assuming that $\overline{a}\alpha > 2$ (which will be verified later), we have that

$$\begin{aligned}
\overline{R}'(t) = \left(\frac{\tau}{t+\tau}\right)^{\overline{a}\alpha}\Bigg(R(0) &+ \overline{b}_1\frac{\alpha^2}{\tau^{\overline{a}\alpha}(\overline{a}\alpha-1)}\left((t+\tau)^{\overline{a}\alpha-1} - \tau^{\overline{a}\alpha-1}\right)\\
&+ 2\overline{b}_2\frac{\alpha^2}{\tau^{\overline{a}\alpha}(\overline{a}\alpha-2)}\left((t+\tau)^{\overline{a}\alpha-2} - \tau^{\overline{a}\alpha-2}\right)\Bigg).
\end{aligned}$$

This readily gives the upper bound

$$\begin{aligned}\overline{R}'(t) &\leq R(0) + \bar{b}_1 \frac{\alpha^2}{(t+\tau)(\bar{a}\alpha - 1)} + 2\bar{b}_2 \frac{\alpha^2}{(t+\tau)^2(\bar{a}\alpha - 2)} \\ &\leq R(0) + \bar{b}_1 \frac{\alpha^2}{\tau(\bar{a}\alpha - 1)} + 2\bar{b}_2 \frac{\alpha^2}{\tau^2(\bar{a}\alpha - 2)}\end{aligned} \tag{J.206}$$

valid for all $t \in [0, 1]$. Furthermore, for $t = 1$, we also obtain the upper bound

$$\begin{aligned}\overline{R}'(1) &\leq \left(\frac{\tau}{1+\tau}\right)^{\bar{a}\alpha} R(0) + \bar{b}_1 \frac{\alpha^2}{(1+\tau)(\bar{a}\alpha - 1)} + 2\bar{b}_2 \frac{\alpha^2}{(1+\tau)^2(\bar{a}\alpha - 2)} \\ &\leq \tau^{\bar{a}\alpha} R(0) + \bar{b}_1 \frac{\alpha^2}{\bar{a}\alpha - 1} + 2\bar{b}_2 \frac{\alpha^2}{\bar{a}\alpha - 2}.\end{aligned} \tag{J.207}$$

Let us now pick

$$\bar{a} = \frac{\lambda_{\min}}{\sqrt{A + 90}}, \qquad \bar{b}_1 = \lambda_{\max} \frac{c^2 \gamma}{2}, \qquad \bar{b}_2 = 4c^2 \frac{\gamma^2}{\rho^2}. \tag{J.208}$$

This implies that

$$\tau = \max\left(\bar{b}_1, 2\sqrt{\bar{b}_2}\right), \qquad \bar{a}\alpha = 3, \tag{J.209}$$

which combined with (J.206) gives

$$\overline{R}'(t) \leq R(0) + \alpha^2 = R(0) + \frac{9(A+90)}{\lambda_{\min}^2}. \tag{J.210}$$

We now verify that, for all $t \in [0, 1]$,

$$\bar{a} \leq 2\lambda_{\min} \mu_c(\overline{R}(t)) = 2\lambda_{\min} \mathrm{erf}\left(\frac{c}{2\sqrt{\overline{R}(t) + \zeta^2/2}}\right). \tag{J.211}$$

For (J.211) to hold, it suffices that

$$\bar{a} \leq 2\lambda_{\min} \mathrm{erf}\left(\frac{c}{2\sqrt{A + 9(A+90)/\lambda_{\min}^2}}\right), \tag{J.212}$$

due to (J.210) and to the fact that $R(0) + \zeta^2/2 \leq A$. Note that

$$\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} \mathrm{d}t \geq \frac{2z}{\sqrt{\pi e}} \geq \frac{2z}{9},$$

where the inequalities holds for all $z \in [0, 1/\sqrt{2}]$. Furthermore, as $c = 3\sqrt{2(A+90)}/\lambda_{\min}$, one can readily verify that

$$\frac{c}{2\sqrt{A + 9(A+90)/\lambda_{\min}^2}} \leq \frac{1}{\sqrt{2}}.$$

Thus, (J.212) is implied by

$$\bar{a} \leq \frac{2c\lambda_{\min}}{9\sqrt{A + 9(A+90)/\lambda_{\min}^2}}, \tag{J.213}$$

which holds due to the definitions of $c, \overline{a}$ and to the fact that $\lambda_{\min} \leq 1$.

By Lemma J.5, we have that

$$\frac{2\nu_c(R)\left(R+\zeta^2/2\right)}{c^2} < 1,$$

which implies that

$$\overline{b}_1 > \lambda_{\max} c^2 \frac{\nu_c(R)\left(\overline{R}(t)+\zeta^2/2\right)}{c^2}\gamma, \tag{J.214}$$

for all $t \in [0,1]$. Now, note that $\mathrm{d}\tilde{\eta}^2(t)/\mathrm{d}t \leq -2\alpha^2/(1+\tau)^3 < 0$ for $t \in [0,1]$. Thus, $\overline{R}'(t) \geq 0$. In fact, if this is not the case, by continuity of $\overline{R}'$, there exists an interval $(t^*, t^*+\delta) \subseteq [0,1]$ s.t. $\overline{R}'(t^*) = 0$ and $\overline{R}'(t) < 0$ for all $t \in (t^*, t^*+\delta)$. However, if $\overline{R}'(t^*) = 0$, then the derivative of $\overline{R}'$ evaluated at t^* is > 0, which is a contradiction.

As a result, we have

$$\overline{f}(t, \overline{R}'(t)) < \overline{f}'(t, \overline{R}'(t)), \qquad \text{for all } \; t \in [0,1]. \tag{J.215}$$

Again, by Lemma J.5, we have that

$$\frac{\mu_c(R)}{c} < \frac{1}{\sqrt{\pi\left(R+\zeta^2/2\right)}}, \qquad \underline{c}_\nu(c,\zeta) < \frac{2\nu_c(R)\left(R+\zeta^2/2\right)}{c^2}.$$

Thus, by setting

$$\underline{a} = 2c\lambda_{\max}\frac{1}{\sqrt{\pi\zeta^2/2}}, \qquad \underline{b}_1 = \frac{c^2\gamma\underline{c}_\nu(c,\zeta)}{2}, \qquad \underline{b}_2 = c^2\frac{\gamma^2}{\rho^2},$$

we have that

$$\underline{f}(t, \underline{R}'(t)) > \underline{f}'(t, \underline{R}'(t)), \qquad \text{for all } \; t \in [0,1]. \tag{J.216}$$

Note that $\overline{f}(t,R)$ and $\underline{f}(t,R)$ are continuous in both variables in the intervals $t \in [0,1]$ and $R \in [0,1]$. Furthermore, they are also Lipschitz in R in these same intervals. Thus, by Theorem 1.3 in Teschl [2012],

$$\underline{R}'(t) < \underline{R}(t), \qquad \overline{R}(t) < \overline{R}'(t), \qquad \text{for all } \; t \in (0,1]. \tag{J.217}$$

Note that

$$\sup_{t\in[0,1]} \tilde{\eta}(t) = \frac{\alpha}{\tau} < \frac{2}{\gamma}, \tag{J.218}$$

which follows from

$$\frac{\alpha}{\tau} \leq \frac{3\sqrt{A+90}}{\lambda_{\min}}\frac{\lambda_{\min}^2}{9\gamma\lambda_{\max}(A+90)} \leq \frac{1}{3\gamma\sqrt{A+90}} \leq \frac{1}{3\gamma\sqrt{90}}. \tag{J.219}$$

Thus, noting that

$$\overline{R}(1) + \frac{c^2\tilde{\eta}^2(1)\gamma^2}{\rho^2} = O_\gamma\left(\frac{\lambda_{\max}}{\lambda_{\min}^4}\gamma + \frac{1}{\lambda_{\min}^4}\frac{\gamma^2}{\rho^2}\right),$$

the desired result follows from Theorem 9.6 and Propositions 9.8 and J.12. □

J.4.4 Proofs on the minimax rates

In this section, we consider the notation $D = (X, Y)$ and $D_i' = (X_i', Y_i')$, with $X, X_i' \in \mathbb{R}^{n\times d}$ and $Y, Y_i' \in \mathbb{R}^n$, where the two neighboring datasets differ in their i-th row (x_i, y_i) and (x_i', y_i'). Assume that all data satisfies Assumption 9.3, i.e., (x_i, y_i) are sampled i.i.d. such that $x_i \sim \mathcal{N}(0, \Sigma)$, $y_i = x_i^\top \theta^* + z_i$, and (x_i', y_i') is sampled from this same distribution independently. In this section, the asymptotic notation will hide only numerical absolute constants, independent of ζ and γ.

Denote by $\mathcal{M}$ a generic $\rho^2/2$-zCDP algorithm. $\mathcal{M}$ applied to D (or D') outputs a random variable in the probability space of the data and the private mechanism. Let $p_{\mathcal{M}(D)}$ ($p_{\mathcal{M}(D_i')}$) be the probability density function of the random variable $\mathcal{M}(D)$ ($\mathcal{M}(D_i')$) when taking into account the randomness of $\mathcal{M}$, X and Y (X_i' and Y_i'). Define L_i to be the random variable

$$L_i = \frac{p_{\mathcal{M}(D)}(\theta')}{p_{\mathcal{M}(D_i')}(\theta')}, \tag{J.220}$$

where θ' is set to be the random variable $\theta' = \mathcal{M}(D_i')$. Then, due to the definition of zCDP (see Definition 9.1), we have that, for all D, D_i' and $\alpha > 1$,

$$\mathbb{E}\left[L_i^\alpha \mid D, D_i'\right] \leq \exp\left(\frac{\alpha(\alpha-1)\rho^2}{2}\right). \tag{J.221}$$

Thus, taking the expectation with respect to D and D_i', we have

$$\mathbb{E}\left[L_i^\alpha\right] \leq \exp\left(\frac{\alpha(\alpha-1)\rho^2}{2}\right). \tag{J.222}$$

As in Cai et al. [2021], let us define the tracing attack

$$A_i(\theta) = z_i x_i^\top (\theta - \theta^*). \tag{J.223}$$

Unless differently stated, in this section, when we use the notation $\mathbb{E}$, we refer to expectations on the randomness of the data (X, Y) and of the algorithm $\mathcal{M}$.

Lemma J.17. *Let Assumption 9.3 hold and let ζ^2 be an absolute positive constant. Then, there exists a prior π on θ^* with support in $\|\theta^*\|_2 < 1$ and an absolute constant $C > 0$ such that, for any algorithm $\mathcal{M}$ with $\sup_{\|\theta^*\|_2<1} \mathbb{E}\left[\|\mathcal{M}\|_2^2\right] < \infty$, either one of the following holds*

$$\mathbb{E}_\pi\left[\sum_{i\in[n]} \mathbb{E}\left[A_i(\mathcal{M}(D))\right]\right] \geq Cd, \qquad \text{or} \qquad \mathbb{E}_{\pi,X,Y,\mathcal{M}}\left[\|\mathcal{M}(D) - \theta^*\|_2^2\right] > C. \tag{J.224}$$

Proof. Consider a generic algorithm $\mathcal{M}$. We have

$$\sum_{i=1}^n \mathbb{E}\left[A_i(\mathcal{M}(D))\right] = \sum_{i=1}^n \mathbb{E}\left[z_i x_i^\top \mathcal{M}(D)\right] = \sum_{j\in[d]} \mathbb{E}\left[\mathcal{M}(D)_j \sum_{i\in[n]} x_{ij}(y_i - x_i^\top \theta^*)\right], \tag{J.225}$$

since the term θ^* vanishes in expectation from the definition of A_i in (J.223) and where we replace $z_i = y_i - x_i^\top \theta^*$ and introduce the notation $x_{ij} \in \mathbb{R}$ to denote the j-th entry of the data vector x_i. By Assumption 9.3, we have

$$p(y_i|x_i, \theta^*) = \mathcal{N}(x_i^\top \theta^*, \zeta^2), \qquad \frac{\partial}{\partial \theta_j^*} p(y_i|x_i, \theta^*) = \frac{x_{ij}(y_i - x_i^\top \theta^*)}{\zeta^2} p(y_i|x_i, \theta^*), \tag{J.226}$$

which give

$$\begin{aligned}\zeta^2 \frac{\partial}{\partial \theta_j^*} \log p(y_i|x_i,\theta^*) &= x_{ij}(y_i - x_i^\top \theta^*),\\ \zeta^2 \frac{\partial}{\partial \theta_j^*} \log p(Y|X,\theta^*) &= \sum_{i\in[n]} x_{ij}(y_i - x_i^\top \theta^*),\end{aligned} \tag{J.227}$$

as $p(Y|X,\theta^*) = \Pi_i p(y_i|x_i,\theta^*)$. Thus, we have (dropping the dependence of $\mathcal{M}$ on X and Y)

$$\begin{aligned}&\mathbb{E}\left[\mathcal{M}_j \sum_{i\in[n]} x_{ij}(y_i - x_i^\top \theta^*)\right] = \mathbb{E}\left[\mathcal{M}_j \zeta^2 \frac{\partial}{\partial \theta_j^*} \log p(Y|X,\theta^*)\right]\\ &= \int \mathcal{M}_j \zeta^2 \left(\frac{1}{p(Y|X,\theta^*)} \frac{\partial}{\partial \theta_j^*} p(Y|X,\theta^*)\right) p(X)p(Y|X,\theta^*)p(\mathcal{M}_j|X,Y)\mathrm{d}X\mathrm{d}Y\,\mathrm{d}\mathcal{M}_j\\ &= \int \mathcal{M}_j \zeta^2 \left(\frac{\partial}{\partial \theta_j^*} p(Y|X,\theta^*)\right) p(X)p(\mathcal{M}_j|X,Y)\mathrm{d}X\mathrm{d}Y\,\mathrm{d}\mathcal{M}_j\\ &= \frac{\partial}{\partial \theta_j^*} \int \mathcal{M}_j \zeta^2 p(X)p(Y|X,\theta^*)p(\mathcal{M}_j|X,Y)\mathrm{d}X\mathrm{d}Y\,\mathrm{d}\mathcal{M}_j\\ &= \zeta^2 \frac{\partial}{\partial \theta_j^*} \mathbb{E}\left[\mathcal{M}_j\right],\end{aligned} \tag{J.228}$$

where the fourth line holds since $p(Y|X)$ is the only term that depends on θ^*, and the derivative is swapped with the integration due to dominated convergence, as in the open ball $\|\theta^*\|_2 < 1$ we uniformly have

$$\mathbb{E}\left[\left(\frac{\partial}{\partial \theta_j^*} \log p(Y|X,\theta^*)\right)^2\right] < \infty, \qquad \mathbb{E}\left[\mathcal{M}_j^2\right] < \infty. \tag{J.229}$$

Here, the first condition follows from Assumption 9.3, and the second one by hypothesis. Notice that the previous equation, by Cauchy-Schwartz inequality, implies the condition for completeness in Theorem 2.1 in Cai et al. [2025], and that this condition also implies the function $g_j(\theta^*) := \mathbb{E}\left[\mathcal{M}_j\right]$ being differentiable for all θ^* in $\|\theta^*\|_2 < 1$. Then, plugging (J.228) in (J.225) and taking the expectation over the prior π of θ^*, we get

$$\mathbb{E}_\pi\left[\sum_{i\in[n]} \mathbb{E}\left[A_i(\mathcal{M}(D))\right]\right] = \zeta^2 \mathbb{E}_\pi\left[\sum_{j\in[d]} \frac{\partial}{\partial \theta_j^*} g_j(\theta^*)\right]. \tag{J.230}$$

Consider the prior π on θ^* such that all the entries θ_j^* are independent and distributed according to

$$\pi_j(z) = \frac{15d^{5/2}}{16}\left(\frac{1}{d} - z^2\right)^2, \tag{J.231}$$

with support in $z \in (-d^{-1/2}, d^{-1/2})$. This prior is such that the support is $\|\theta^*\|_2 < 1$, it vanishes at the boundary and

$$-\mathbb{E}_{z\sim\pi_j}\left[-z\frac{\pi_j'(z)}{\pi_j(z)}\right] = 1, \qquad \mathbb{E}_{z\sim\pi_j}\left[\left(\frac{\pi_j'(z)}{\pi_j(z)}\right)^2\right] = 10d. \tag{J.232}$$

Let us define the one dimensional function

$$h_j(\theta_j^*) := \mathbb{E}_{\theta^*_{-j}}\left[g_j(\theta^*)\right], \tag{J.233}$$

where the notation $\mathbb{E}_{\theta^*_{-j}}$ takes expectation over all the (independent) entries of θ^*, except for the j-th. Then, due to (J.229), we can apply Stein's Lemma as stated in Lemma 2.1 in Cai et al. [2025], which yields

$$\mathbb{E}_\pi\left[\frac{\partial}{\partial\theta_j^*}g_j(\theta^*)\right] = \mathbb{E}_{\theta_j^*}\left[\frac{\partial}{\partial\theta_j^*}h_j(\theta_j^*)\right] = \mathbb{E}_{\theta_j^*}\left[-h_j(\theta_j^*)\frac{\pi_j'(\theta_j^*)}{\pi_j(\theta_j^*)}\right] = \mathbb{E}_\pi\left[-g_j(\theta^*)\frac{\pi_j'(\theta_j^*)}{\pi_j(\theta_j^*)}\right]. \tag{J.234}$$

By triangle inequality, we have

$$-g_j(\theta^*)\frac{\pi_j'(\theta_j^*)}{\pi_j(\theta_j^*)} \geq -\theta_j^*\frac{\pi_j'(\theta_j^*)}{\pi_j(\theta_j^*)} - \left|(g_j(\theta^*)-\theta_j^*)\frac{\pi_j'(\theta_j^*)}{\pi_j(\theta_j^*)}\right|. \tag{J.235}$$

Plugging in (J.234) and using (J.232), we get

$$\mathbb{E}_\pi\left[\frac{\partial}{\partial\theta_j^*}g_j(\theta^*)\right] \geq 1 - \mathbb{E}_\pi\left[\left|(g_j(\theta^*)-\theta_j^*)\frac{\pi_j'(\theta_j^*)}{\pi_j(\theta_j^*)}\right|\right]. \tag{J.236}$$

Summing over j we get

$$\begin{aligned}
\sum_{j\in[d]}\mathbb{E}_\pi\left[\frac{\partial}{\partial\theta_j^*}g_j(\theta^*)\right] &\geq d - \mathbb{E}_\pi\left[\sum_{j\in[d]}\left|(g_j(\theta^*)-\theta_j^*)\frac{\pi_j'(\theta_j^*)}{\pi_j(\theta_j^*)}\right|\right] \\
&\geq d - \mathbb{E}_\pi\left[\sqrt{\sum_{j\in[d]}\left(g_j(\theta^*)-\theta_j^*\right)^2}\sqrt{\sum_{j\in[d]}\left(\frac{\pi_j'(\theta_j^*)}{\pi_j(\theta_j^*)}\right)^2}\right] \\
&\geq d - \sqrt{\mathbb{E}_\pi\left[\sum_{j\in[d]}\left(g_j(\theta^*)-\theta_j^*\right)^2\right]}\sqrt{\mathbb{E}_\pi\left[\sum_{j\in[d]}\left(\frac{\pi_j'(\theta_j^*)}{\pi_j(\theta_j^*)}\right)^2\right]} \\
&\geq d - \sqrt{\mathbb{E}_\pi\left[\left\|\mathbb{E}\left[\mathcal{M}(D)\right]-\theta^*\right\|_2^2\right]}\sqrt{10d^2} \\
&\geq d\left(1 - \sqrt{10\,\mathbb{E}_{\pi,X,Y,\mathcal{M}}\left[\left\|\mathcal{M}(D)-\theta^*\right\|_2^2\right]}\right).
\end{aligned} \tag{J.237}$$

Plugging this inequality in (J.230) we get

$$\mathbb{E}_\pi\left[\sum_{i\in[n]}\mathbb{E}\left[A_i(\mathcal{M}(D))\right]\right] \geq \zeta^2 d\left(1 - \sqrt{10\,\mathbb{E}_{\pi,X,Y,\mathcal{M}}\left[\left\|\mathcal{M}(D)-\theta^*\right\|_2^2\right]}\right). \tag{J.238}$$

Suppose $\mathbb{E}_{\pi,X,Y,\mathcal{M}}\left[\left\|\mathcal{M}(D)-\theta^*\right\|_2^2\right] \leq 1/40$. Then, we have $\mathbb{E}_\pi\left[\sum_{i\in[n]}\mathbb{E}\left[A_i(\mathcal{M}(D))\right]\right] \geq \zeta^2 d/2$. Thus, the desired result follows by taking $C = \min(\zeta^2/2, 1/40)$. □

Lemma J.18. *Let Assumption 9.3 hold, ζ^2 be an absolute positive constant and $d^2/(\rho^2 n^2 \lambda_{\min}) = O(1)$. Define $\mathbb{M}$ to be the set of all $\rho^2/2$-zCDP algorithms such that $\sup_{\|\theta^*\|_2<1}\mathbb{E}\left[\|\mathcal{M}\|_2^2\right] < \infty$. Then, there exists a prior π on θ^* with support in $\|\theta^*\|_2 < 1$ such that*

$$\inf_{\mathcal{M}\in\mathbb{M}}\mathbb{E}_{\pi,\mathcal{M},X,Y}\left[\mathcal{R}(\mathcal{M}(X,Y))\right] = \Omega\left(\frac{d^2}{n^2\left(e^{\rho^2}-1\right)}\right). \tag{J.239}$$

Proof. Note that, according to (J.223) and (J.220), we have

$$\mathbb{E}\left[A_i(\mathcal{M}(D_i'))\right] = 0, \qquad \mathbb{E}\left[(L_i - 1)A_i(\mathcal{M}(D_i'))\right] = \mathbb{E}\left[L_i A_i(\mathcal{M}(D_i'))\right] = \mathbb{E}\left[A_i(\mathcal{M}(D))\right]. \tag{J.240}$$

The first one holds because z_i, x_i and $(\mathcal{M}(D_i') - \theta^*)$ are independent, and the first one is mean-0. The last equality follows from

$$\begin{aligned}\mathbb{E}\left[L_i A_i(\mathcal{M}(D_i'))\right] &= \int \frac{p_{\mathcal{M}(D)}(\theta)}{p_{\mathcal{M}(D_i')}(\theta)} A_i(\theta) p_{\mathcal{M}(D_i')}(\theta) \mathsf{d}\theta \\ &= \int A_i(\theta) p_{\mathcal{M}(D)}(\theta) \mathsf{d}\theta \\ &= \mathbb{E}\left[A_i(\mathcal{M}(D))\right].\end{aligned} \tag{J.241}$$

Then, we have

$$\begin{aligned}\sum_{i=1}^n \mathbb{E}\left[A_i(\mathcal{M}(D))\right] &= \sum_{i=1}^n \mathbb{E}\left[(L_i - 1)A_i(\mathcal{M}(D_i'))\right] \\ &\le \sum_{i=1}^n \sqrt{\mathbb{E}\left[(L_i - 1)^2\right]} \sqrt{\mathbb{E}\left[A_i(\mathcal{M}(D_i'))^2\right]} \\ &\le n\sqrt{\mathbb{E}\left[(L_1 - 1)^2\right]} \sqrt{\mathbb{E}\left[A_1(\mathcal{M}(D_1'))^2\right]},\end{aligned} \tag{J.242}$$

where we used Cauchy-Schwartz in the second line and the fact that the samples are equally distributed in the last line.

Since $\mathbb{E}[L_1] = 1$, we have

$$\begin{aligned}\mathbb{E}\left[(L_1 - 1)^2\right] &= \mathbb{E}\left[L_1^2\right] - 1 \le \exp\left(\rho^2\right) - 1, \\ \mathbb{E}\left[A_1(\mathcal{M}(D_1'))^2\right] &= \mathbb{E}\left[z_1^2 \left(x_1^\top \left(\mathcal{M}(D_1') - \theta^*\right)\right)^2\right] \\ &= \zeta^2 \mathbb{E}\left[\left\|\Sigma^{1/2}\left(\mathcal{M}(D_1') - \theta^*\right)\right\|_2^2\right] \\ &= 2\zeta^2 \mathbb{E}\left[\mathcal{R}\left(\mathcal{M}(D)\right)\right],\end{aligned} \tag{J.243}$$

where the steps in the third line are again motivated by the fact that z_1, x_1 and $(\mathcal{M}(D_1') - \theta^*)$ are independent.

Let us now focus on the LHS of (J.242). Suppose

$$\mathbb{E}_{\pi,\mathcal{M},X,Y}\left[\left\|\mathcal{M}(D) - \theta^*\right\|_2^2\right] \le \frac{2\,\mathbb{E}_{\pi,\mathcal{M},X,Y}\left[\mathcal{R}(\mathcal{M}(D))\right]}{\lambda_{\min}} < C_1. \tag{J.244}$$

Setting C_1 to be the same constant as in Lemma J.17, by (J.224) we have that

$$\mathbb{E}_\pi\left[\sum_{i\in[n]} \mathbb{E}\left[A_i(\mathcal{M}(D))\right]\right] \ge C_1 d. \tag{J.245}$$

This, together with (J.243) and (J.242), after taking the expectation with respect to π yields

$$\mathbb{E}_{\pi,\mathcal{M},X,Y}\left[\mathcal{R}\left(\mathcal{M}(D)\right)\right] = \Omega\left(\frac{d^2}{n^2\left(e^{\rho^2} - 1\right)}\right), \tag{J.246}$$

which yields the desired result.

Suppose instead that

$$\frac{2\,\mathbb{E}_{\pi,\mathcal{M},X,Y}\left[\mathcal{R}(\mathcal{M}(D))\right]}{\lambda_{\min}} \geq C_1. \tag{J.247}$$

Then, by hypothesis, we would have

$$\mathbb{E}_{\pi,\mathcal{M},X,Y}\left[\mathcal{R}(\mathcal{M}(D))\right] \geq \frac{\lambda_{\min}C_1}{2} = \Omega\left(\frac{d^2}{n^2\rho^2}\right) = \Omega\left(\frac{d^2}{n^2\left(e^{\rho^2}-1\right)}\right), \tag{J.248}$$

which again yields the desired result.

□

Lemma J.19. *Let Assumption 9.3 hold, ζ^2 be an absolute positive constant, $d < n = O(d^2)$, $\lambda_{\min} > 0$ and $\lambda_{\max} = O(1)$. Define $\mathbb{M}$ to be the set of all algorithms such that $\sup_{\|\theta^*\|_2<1}\mathbb{E}\left[\|\mathcal{M}\|_2^2\right] < \infty$. Then, we have*

$$\inf_{\mathcal{M}\in\mathbb{M}}\sup_{\|\theta^*\|_2<1}\mathbb{E}\left[\mathcal{R}(\mathcal{M}(X,Y))\right] = \Omega\left(\frac{d}{n}\right). \tag{J.249}$$

Proof. Let us denote π_B as the Gaussian distribution in d dimensions $\mathcal{N}(0,\zeta^2/(\lambda n)I)$, with $\lambda = 2\lambda_{\max}$, truncated at $\|\theta^*\|_2 \geq 1$. Then, we have

$$\inf_{\mathcal{M}\in\mathbb{M}}\sup_{\|\theta^*\|_2<1}\mathbb{E}\left[\mathcal{R}(\mathcal{M}(X,Y))\right] \geq \inf_{\mathcal{M}\in\mathbb{M}}\mathbb{E}_{\theta^*\sim\pi_B}\mathbb{E}\left[\mathcal{R}(\mathcal{M}(X,Y))\right]. \tag{J.250}$$

Let B be the closed unit ball, and let $\mathcal{B}$ be the ellipsoid $\Sigma^{1/2}B$. Given any algorithm $\mathcal{M}$, define the composition $\mathcal{M}_B = \Sigma^{-1/2}\operatorname{Proj}_{\mathcal{B}}\Sigma^{1/2}\circ\mathcal{M}$. This implies that, for all (X,Y), $\mathcal{M}_B(X,Y)\in B$. Furthermore, for all (X,Y) and $\theta^*\in B$, we have

$$\begin{aligned}\left\|\Sigma^{1/2}(\mathcal{M}_B(X,Y)-\theta^*)\right\|_2^2 &= \left\|\operatorname{Proj}_{\mathcal{B}}\Sigma^{1/2}\mathcal{M}(X,Y)-\Sigma^{1/2}\theta^*\right\|_2^2\\ &= \left\|\operatorname{Proj}_{\mathcal{B}}\Sigma^{1/2}\mathcal{M}(X,Y)-\operatorname{Proj}_{\mathcal{B}}\Sigma^{1/2}\theta^*\right\|_2^2\\ &\leq \left\|\Sigma^{1/2}\left(\mathcal{M}(X,Y)-\theta^*\right)\right\|_2^2,\end{aligned} \tag{J.251}$$

where the second step holds since $\Sigma^{1/2}\theta^*\in\mathcal{B}$, and the last step holds since projections on closed convex sets are non-expansive. Thus, we have $\mathcal{R}(\mathcal{M}_B(X,Y)) \leq \mathcal{R}(\mathcal{M}(X,Y))$. Then, since the support of π_B is included in B, it holds that

$$\inf_{\mathcal{M}\in\mathbb{M}}\mathbb{E}_{\theta^*\sim\pi_B}\mathbb{E}\left[\mathcal{R}(\mathcal{M}(X,Y))\right] = \inf_{\mathcal{M}\in\mathbb{M}}\mathbb{E}_{\theta^*\sim\pi_B}\mathbb{E}\left[\mathcal{R}(\mathcal{M}_B(X,Y))\right], \tag{J.252}$$

where the equality follows from the fact that the set of algorithms $\{\mathcal{M}_B \,|\, \mathcal{M}\in\mathbb{M}\}$ is a subset of $\mathbb{M}$.

Let us denote with $\pi = \mathcal{N}(0,\zeta^2/(\lambda n)I)$ the non-truncated Gaussian distribution and with A the event that $\theta^*\sim\pi$ is such that $\|\theta^*\|_2 < 1$. Then, dropping the notation $\mathbb{M}$ and the dependence of $\mathcal{M}$ from (X,Y), one has

$$\begin{aligned}\inf_{\mathcal{M}}\mathbb{E}_{\theta^*\sim\pi_B}\mathbb{E}\left[\mathcal{R}(\mathcal{M}_B)\right] &= \frac{1}{\mathbb{P}(A)}\inf_{\mathcal{M}}\mathbb{E}_{\theta^*\sim\pi}\mathbb{E}\left[\mathcal{R}(\mathcal{M}_B)\mathbf{1}_A\right]\\ &= \frac{1}{\mathbb{P}(A)}\inf_{\mathcal{M}}\mathbb{E}_{\theta^*\sim\pi}\mathbb{E}\left[\mathcal{R}(\mathcal{M}_B)(1-\mathbf{1}_{A^c})\right]\\ &\geq \frac{1}{\mathbb{P}(A)}\inf_{\mathcal{M}}\mathbb{E}_{\theta^*\sim\pi}\mathbb{E}\left[\mathcal{R}(\mathcal{M}_B)\right] - \frac{1}{\mathbb{P}(A)}\sup_{\mathcal{M}}\mathbb{E}_{\theta^*\sim\pi}\mathbb{E}\left[\mathcal{R}(\mathcal{M}_B)\mathbf{1}_{A^c}\right].\end{aligned} \tag{J.253}$$

Note that

$$\sup_{\mathcal{M}} \mathbb{E}_{\theta^* \sim \pi} \mathbb{E}\left[\mathcal{R}(\mathcal{M}_B) \mathbf{1}_{A^c}\right] \leq \sup_{\mathcal{M}} \sqrt{\mathbb{E}_{\theta^* \sim \pi} \mathbb{E}\left[\mathcal{R}(\mathcal{M}_B)^2\right]} \sqrt{\mathbb{P}(A^c)} \leq 2\lambda_{\max} \sqrt{\mathbb{P}(A^c)}, \quad \text{(J.254)}$$

where the first step follows from Cauchy-Schwartz inequality and the second step holds since $\|\mathcal{M}_B\|_2 \leq 1$. Furthermore, since $\lambda > 2$ and $n > d$, due to concentration of the norm of Gaussian vectors we have that $\mathbb{P}(A) \geq 1 - 2e^{-c_1 d}$ for some absolute constant c_1. Then, we also have

$$\begin{aligned}
\inf_{\mathcal{M}} \mathbb{E}_{\theta^* \sim \pi} \mathbb{E}\left[\mathcal{R}(\mathcal{M}_B)\right] &\geq \inf_{\mathcal{M}} \mathbb{E}_{\theta^* \sim \pi} \mathbb{E}\left[\mathcal{R}(\mathcal{M})\right] \\
&= \frac{\zeta^2}{n} \mathbb{E}\left[\operatorname{tr}\left((X^\top X/n + \lambda I)^{-1} \Sigma\right)\right] \\
&\geq \frac{\zeta^2}{n} \operatorname{tr}\left(\left(\mathbb{E}\left[\Sigma^{-1/2}(X^\top X/n + \lambda I)\Sigma^{-1/2}\right]\right)^{-1}\right) \\
&= \frac{\zeta^2}{n} \operatorname{tr}\left(\left(I + \lambda \Sigma^{-1}\right)^{-1}\right) \\
&\geq \frac{\zeta^2}{2\lambda n} \operatorname{tr}\left(\Sigma\right) \\
&= \frac{\zeta^2 d}{4\lambda_{\max} n},
\end{aligned} \quad \text{(J.255)}$$

where the second step follows from the closed form of the Bayes estimator under Gaussian prior (see, e.g., Eq. (63) in Mourtada [2022]), the third step follows from Jensen inequality since the mapping $A \mapsto \operatorname{tr}(A^{-1})$ is convex if A is positive definite (see Eq. (1.33) in Bhatia [2007]), and the last two steps follow from the fact that λ has been set to $2\lambda_{\max}$ and $\operatorname{tr}(\Sigma) = d$.

Then, putting (J.253), (J.254) and (J.255) together, we get

$$\inf_{\mathcal{M}} \mathbb{E}_{\theta^* \sim \pi_B} \mathbb{E}\left[\mathcal{R}(\mathcal{M}_B)\right] \geq \frac{\zeta^2 d}{4\lambda_{\max} n} - 4\lambda_{\max} e^{-c_1 d/2}. \quad \text{(J.256)}$$

Since $n = O(d^2)$, merging with (J.250), (J.252) yields the desired result. □

Proof of Theorem 9.12. Note that, if $\sup_{\|\theta^*\|_2 < 1} \mathbb{E}\left[\|\mathcal{M}\|_2^2\right]$ diverges, then $\sup_{\|\theta^*\|_2 < 1} \mathbb{E}\left[\mathcal{R}(\mathcal{M}(X, Y))\right]$ diverges as well. Thus, let us focus on the class of $\rho^2/2$-zCDP algorithms such that $\sup_{\|\theta^*\|_2 < 1} \mathbb{E}\left[\|\mathcal{M}\|_2^2\right] < \infty$. By Lemmas J.18 and J.19, and using the same argument as in (J.250), we have

$$\inf_{\mathcal{M} \in \mathbb{M}} \sup_{\|\theta^*\|_2 < 1} \mathbb{E}\left[\mathcal{R}(\mathcal{M}(X, Y))\right] = \Omega\left(\frac{d}{n} + \frac{d^2}{(e^{\rho^2} - 1)n^2}\right). \quad \text{(J.257)}$$

Notice that, if $\rho > 1$,

$$\frac{d^2}{(e^{\rho^2} - 1)n^2} = O\left(\frac{d}{n}\right), \qquad \frac{d^2}{\rho^2 n^2} = O\left(\frac{d}{n}\right), \quad \text{(J.258)}$$

i.e., the first terms in the RHS of (J.260) and (J.257) dominate. On the other hand, if $\rho < 1$, we have

$$\frac{d^2}{(e^{\rho^2} - 1)n^2} = \Theta\left(\frac{d^2}{\rho^2 n^2}\right). \quad \text{(J.259)}$$

Thus, for any value of $\rho > 0$, we have that

$$\inf_{\mathcal{M}\in\mathbb{M}} \sup_{\|\theta^*\|_2<1} \mathbb{E}\left[\mathcal{R}(\mathcal{M}(X,Y))\right] = \Omega\left(\frac{d}{n} + \frac{d^2}{n^2\rho^2}\right), \tag{J.260}$$

which gives the desired result when taking sequentially the limits $d/n \to \gamma$ and $\gamma \to 0$. □

J.5 Proofs for Section 9.6

In this section, we always assume Assumptions 9.3, 9.4 and 9.13 to hold. We consider the limit $\gamma \to 0$, and use the notation $f(\gamma) \asymp g(\gamma)$ to indicate two quantities such that $f(\gamma) = \Theta_\gamma(g(\gamma))$, where the notation Θ_γ does not track constants depending on ϕ, ψ, ζ, $\|\theta^*\|_2^2$, and α. Similarly, we will consider the notations $\gtrsim$ and $\lesssim$ for $\Omega_\gamma(\cdot)$ and $O_\gamma(\cdot)$ respectively. Furthermore, unless differently specified, we will always consider schedules such that $\tilde{\eta}(t) \leq 2/\gamma$ uniformly in t, so that $\bar{\eta}(t) = \tilde{\eta}(t)$.

We have

$$\Gamma(t) = \int_0^t \tilde{\eta}(s)\mu_c(R(s))\mathsf{d}s, \qquad F(t) := \int_0^t \tilde{\eta}(s)\mathsf{d}s, \tag{J.261}$$

where we defined the shorthand $F(t)$. Then, (9.63) reads

$$\begin{aligned} R(t) &= \mathcal{F}(\Gamma(t)) + \\ &+ \int_0^t \left(\tilde{\eta}^2(s)\nu_c(R(s))(R(s)+\zeta^2/2)\gamma\mathcal{K}(\Gamma(t)-\Gamma(s)) + 2c^2\tilde{\sigma}^2(s)\gamma^2\mathcal{J}(\Gamma(t)-\Gamma(s))\right)\mathsf{d}s. \end{aligned} \tag{J.262}$$

Within this section, for convenience, we also introduce the shorthand $v = c\tilde{\eta}(0)$.

Proposition J.20. *Let $\mathcal{F}(x)$, $\mathcal{K}(x)$, $\mathcal{J}(x)$ be defined according to* (9.64)*. Consider the shorthands*

$$\kappa_1 = 2 - \phi - \psi, \qquad \kappa_2 = 3 - \phi, \qquad \kappa_3 = 2 - \phi, \tag{J.263}$$

so that $\kappa_1 > 1$, $\kappa_2 > 2$, and $\kappa_3 > 1$ due to Assumption 9.13. Then, we have that $\mathcal{F}(x) \asymp x^{-\kappa_1}$, $\mathcal{K}(x) \asymp x^{-\kappa_2}$, and $\mathcal{J}(x) \asymp x^{-\kappa_3}$ uniformly for $x \geq 1$, and $\mathcal{F}(x) \asymp \mathcal{K}(x) \asymp \mathcal{J}(x) \asymp 1$ uniformly for $x < 1$.

Proof. Due to Assumption 9.13, we have that, for $d \to \infty$,

$$\mathcal{F}(x) \to \int p(\lambda)\lambda^{-\psi+1}e^{-2\lambda x}\mathsf{d}\lambda, \qquad \mathcal{K}(x) \to \int p(\lambda)\lambda^2 e^{-2\lambda x}\mathsf{d}\lambda, \qquad \mathcal{J}(x) \to \int p(\lambda)\lambda e^{-2\lambda x}\mathsf{d}\lambda. \tag{J.264}$$

Then, for the second part of the statement, notice that $\mathcal{F}(x)$, $\mathcal{K}(x)$ and $\mathcal{J}(x)$ are decreasing in x, and a direct computation for $x \in \{0,1\}$ yields the desired result. For $x \geq 1$, the desired result is given by the same approach used in Lemma D.5 in Collins-Woodfin et al. [2024b]. □

Lemma J.21. *Assume $c \lesssim 1$ and $\tilde{\eta}(t) \leq 2/\gamma$ uniformly in $t \in [0,1]$. Then, we have*

$$\begin{aligned} R(t) &\asymp \mathcal{F}(\Gamma(t)) + \int_0^t \left(\tilde{\eta}^2(s)c^2\gamma\mathcal{K}(\Gamma(t)-\Gamma(s)) + c^2\tilde{\sigma}^2(s)\gamma^2\mathcal{J}(\Gamma(t)-\Gamma(s))\right)\mathsf{d}s \\ &\gtrsim \mathcal{F}(cF(t)) + \int_0^t \left(\tilde{\eta}^2(s)c^2\gamma\mathcal{K}\left(cF(t)-cF(s)\right) + c^2\tilde{\sigma}^2(s)\gamma^2\mathcal{J}\left(cF(t)-cF(s)\right)\right)\mathsf{d}s \\ &\gtrsim \mathcal{F}(cF(t)). \end{aligned} \tag{J.265}$$

Furthermore, if $R(t)$ is uniformly upper bounded by a constant not dependent on γ, we have

$$R(t) \asymp \mathcal{F}(cF(t)) + \int_0^t \left(\tilde{\eta}^2(s)c^2\gamma\mathcal{K}\left(cF(t) - cF(s)\right) + c^2\tilde{\sigma}^2(s)\gamma^2\mathcal{J}\left(cF(t) - cF(s)\right)\right)\mathrm{d}s. \tag{J.266}$$

Proof. By Lemma J.5, we have that $\mu_c(R) \lesssim c$. Then, multiplying both sides by $\tilde{\eta}(s)$ and integrating from s to t, we get

$$\Gamma(t) - \Gamma(s) \lesssim c\left(F(t) - F(s)\right), \tag{J.267}$$

and $\Gamma(t) \lesssim cF(t)$ when setting $s = 0$. Then, since $\mathcal{F}$, $\mathcal{K}$ and $\mathcal{J}$ are all uniformly $\asymp$ to decreasing functions, we have

$$\begin{aligned} \mathcal{F}(\Gamma(t)) &\gtrsim \mathcal{F}(cF(t)) \\ \mathcal{K}\left(\Gamma(t) - \Gamma(s)\right) &\gtrsim \mathcal{K}\left(cF(t) - cF(s)\right) \\ \mathcal{J}\left(\Gamma(t) - \Gamma(s)\right) &\gtrsim \mathcal{J}\left(cF(t) - cF(s)\right). \end{aligned} \tag{J.268}$$

Furthermore, again due to Lemma J.5, since $c \lesssim 1$, we have $c^2 \asymp \nu_c(R(s))(R(s) + \zeta^2/2)$. Thus, since all terms on the RHS of (J.262) are positive, and dropping numerical constants, we get the first part of the thesis.

For the second part of the thesis, notice that Lemma J.5 gives

$$\frac{c}{\sqrt{\zeta^2 + 2R}} \lesssim \mu_c(R) \lesssim c, \tag{J.269}$$

which implies that, if $\sup_t R(t) \lesssim 1$, the relations $\lesssim$ in (J.267) can be replaced by $\asymp$. Thus, the thesis follows from Proposition J.20, since we have $\mathcal{F}(x) \asymp \mathcal{F}(ax)$ if $a \asymp 1$, with the same relation holding for $\mathcal{K}$ and $\mathcal{J}$. □

Lemma J.22. *Assume $c \lesssim 1$ and $\tilde{\eta}(t) \leq 2/\gamma$ uniformly in $t \in [0, 1]$. Then, if*

$$\mathcal{F}(cF(t)) + \int_0^t \left(\tilde{\eta}^2(s)c^2\gamma\mathcal{K}\left(cF(t) - cF(s)\right) + c^2\tilde{\sigma}^2(s)\gamma^2\mathcal{J}\left(cF(t) - cF(s)\right)\right)\mathrm{d}s \lesssim 1, \tag{J.270}$$

uniformly in $t \in [0, 1]$, we have that

$$R(t) \asymp \mathcal{F}(cF(t)) + \int_0^t \left(\tilde{\eta}^2(s)c^2\gamma\mathcal{K}\left(cF(t) - cF(s)\right) + c^2\tilde{\sigma}^2(s)\gamma^2\mathcal{J}\left(cF(t) - cF(s)\right)\right)\mathrm{d}s. \tag{J.271}$$

Proof. Define the following system of ODEs

$$\mathrm{d}D_i'(t) = -\lambda_i\tilde{\eta}(t)cD_i'(t)\mathrm{d}t + \lambda_i\tilde{\eta}^2(t)c^2\gamma\mathrm{d}t + c^2\tilde{\sigma}^2(t)\gamma^2\mathrm{d}t, \tag{J.272}$$

with $D_i'(0) = D_i(0) = d\left(\omega_i^\top\theta^*\right)^2/2$, and $R'(t) = \frac{1}{d}\sum_{i=1}^d \lambda_i D_i'(t)$. Then, following the same steps as in (9.62) and (9.63) we have that $R'(t)$ is equal to the LHS in (J.270).

By hypothesis, there exists a large enough constant C_1 such that $R'(t) \leq C_1$ for all $t \in [0, 1]$. Then, consider the new system of ODEs

$$\mathrm{d}\underline{D}_i'(t) = -\lambda_i\tilde{\eta}(t)\sqrt{2C_1 + \zeta^2}\frac{c}{\sqrt{2\underline{R}'(t) + \zeta^2}}\underline{D}_i'(t)\mathrm{d}t + \lambda_i\tilde{\eta}^2(t)c^2\gamma\mathrm{d}t + c^2\tilde{\sigma}^2(t)\gamma^2\mathrm{d}t, \tag{J.273}$$

with $\underline{D}_i'(0) = D_i'(0)$, and $\underline{R}'(t) = \frac{1}{d}\sum_{i=1}^d \lambda_i \underline{D}_i'(t)$. Then, by ODE comparison, we have $\underline{R}'(t) \le R'(t)$ for all $t \in [0,1]$. This follows from the same argument as in Theorem 1.3 in Teschl [2012], since $R'(t)$ ($\underline{R}'(t)$) is non-decreasing in $D_i'(t)$ ($\underline{D}_i'(t)$) for all i.

Due to Lemma J.5, since $c \lesssim 1$, we have that there exists a sufficiently large constant C_2 such that $\mu_c(R)C_2 \ge c/\sqrt{2R+\zeta^2}$. Then, by ODE comparison, integrating as in (9.62) and (9.63), and denoting $C_3 = C_2\sqrt{2C_1+\zeta^2}$, we get

$$\begin{aligned}\underline{R}'(t) &\ge \mathcal{F}(C_3\Gamma(t)) + \int_0^t \left(\tilde{\eta}^2(s)c^2\gamma\mathcal{K}\left(C_3\Gamma(t) - C_3\Gamma(s)\right) + c^2\tilde{\sigma}^2(s)\gamma^2\mathcal{J}\left(C_3\Gamma(t) - C_3\Gamma(s)\right)\right)\mathrm{d}s \\ &\asymp R(t),\end{aligned} \tag{J.274}$$

where the last step holds since $c^2 \asymp \nu_c(R(s))(R(s)+\zeta^2/2)$ (due to Lemma J.5 since $c \lesssim 1$), and since $\mathcal{F}(x) \asymp \mathcal{F}(ax)$ if $a \asymp 1$ (with the same relation holding for $\mathcal{K}$ and $\mathcal{J}$) due to Proposition J.20.

Thus, we have

$$R(t) \gtrsim R'(t) \ge \underline{R}'(t) \gtrsim R(t), \tag{J.275}$$

where the first inequality is a consequence of Lemma J.21, which directly yields $R(t) \asymp R'(t)$, proving the thesis. □

Lemma J.23. *Let $\tilde{\eta}(t) = \tilde{\eta}(0)(1-t)^\alpha$, with $\alpha \in \{0, 1/2\}$ or $\alpha \ge 1$. Assume $c \lesssim 1$, $\tilde{\eta}(0) \le 2/\gamma$, $v = \tilde{\eta}(0)c = \omega_\gamma(1)$. Then, we have that*

$$R(1) + \frac{2c^2\tilde{\eta}^2(1)\gamma^2}{\rho^2} \gtrsim v^{-\kappa_1} + \gamma v^{\frac{1}{\alpha+1}} + \frac{\gamma^2}{\rho^2}g(v), \tag{J.276}$$

where we introduced

$$g(v) = \begin{cases} v^{\frac{2}{\alpha+1}} & \text{if } \phi < \dfrac{2}{\alpha+1}, \\ v^\phi & \text{if } \phi > \dfrac{2}{\alpha+1}, \\ v^\phi \ln v & \text{if } \phi = \dfrac{2}{\alpha+1}. \end{cases} \tag{J.277}$$

Furthermore, the relation $\gtrsim$ in (J.276) *becomes $\asymp$ in the following cases:* (i) $\alpha = 0$, *or* (ii) $\alpha = 1/2$ *and* $\gamma^2 v^{4/3}/\rho^2 \lesssim 1$, *or* (iii) $\alpha \ge 1$ *and* $\gamma^2 v/\rho^2 \lesssim 1$.

Proof. By Lemma J.21, we have that

$$R(t) \gtrsim \mathcal{F}(cF(t)) + \int_0^t \left(\tilde{\eta}^2(s)c^2\gamma\mathcal{K}\left(cF(t) - cF(s)\right) + c^2\tilde{\sigma}^2(s)\gamma^2\mathcal{J}\left(cF(t) - cF(s)\right)\right)\mathrm{d}s, \tag{J.278}$$

where $\gtrsim$ can be replaced by $\asymp$ if the RHS is shown to be $\lesssim 1$ for all $t \in [0,1]$, due to Lemma J.22.

Let us focus on the first term of the RHS of (J.278). Due to (9.33), (J.261) gives

$$cF(t) = \frac{v}{(\alpha+1)}(1 - (1-t)^{\alpha+1}). \tag{J.279}$$

Then, due to Proposition J.20, we have

$$\mathcal{F}(cF(1)) \asymp v^{-\kappa_1}, \tag{J.280}$$

with $\mathcal{F}(cF(t)) \lesssim 1$ for all $t \in [0,1]$.

Let us now focus on the the second term of the RHS of (J.278), i.e.,

$$N_2(t) := \gamma c^2 \int_0^t \tilde{\eta}^2(s)\mathcal{K}(cF(t) - cF(s))\mathsf{d}s. \tag{J.281}$$

Consider the change of variable $u = cF(t) - cF(s)$. This yields

$$N_2(t) = \gamma c \int_0^{cF(t)} \tilde{\eta}(s(u))\mathcal{K}(u)\mathsf{d}u. \tag{J.282}$$

Furthermore, (J.279) yields

$$u = \frac{v}{\alpha+1}\left((1-s)^{\alpha+1} - (1-t)^{\alpha+1}\right), \qquad s = 1 - \left(\frac{u(\alpha+1)}{v} + (1-t)^{\alpha+1}\right)^{\frac{1}{\alpha+1}}. \tag{J.283}$$

Then, plugging the last relation in (9.33) we get

$$c\tilde{\eta}(s(u)) = v\left(\frac{u(\alpha+1)}{v} + (1-t)^{\alpha+1}\right)^{\frac{\alpha}{\alpha+1}}. \tag{J.284}$$

Plugging this in (J.282) we obtain

$$N_2(t) = \gamma v^{\frac{1}{\alpha+1}} \int_0^{cF(t)} \left(u(\alpha+1) + v(1-t)^{\alpha+1}\right)^{\frac{\alpha}{\alpha+1}} \mathcal{K}(u)\mathsf{d}u. \tag{J.285}$$

Using Proposition J.20, and the fact that $cF(1) \asymp v = \omega_\gamma(1)$, we have

$$\begin{aligned} N_2(1) &\asymp \gamma v^{\frac{1}{\alpha+1}} \left(\int_0^1 u^{\frac{\alpha}{\alpha+1}}\mathcal{K}(u)\mathsf{d}u + \int_1^{cF(1)} u^{\frac{\alpha}{\alpha+1}}\mathcal{K}(u)\mathsf{d}u\right) \\ &\asymp \gamma v^{\frac{1}{\alpha+1}} \left(\int_0^1 u^{\frac{\alpha}{\alpha+1}}\mathsf{d}u + \int_1^{cF(1)} u^{\frac{\alpha}{\alpha+1}}u^{-\kappa_2}\mathsf{d}u\right) \\ &\asymp \gamma v^{\frac{1}{\alpha+1}}, \end{aligned} \tag{J.286}$$

where the last line holds since $\frac{\alpha}{\alpha+1} - \kappa_2 < -1$ for any $\alpha \geq 0$, since $\kappa_2 > 2$. Notice that, with a similar computation, we also have

$$N_2(t) \lesssim \gamma v \lesssim 1, \tag{J.287}$$

uniformly in $t \in [0,1]$, since $c \lesssim 1$ and $\tilde{\eta}(0) \leq 2/\gamma$ by hypothesis.

Let us now focus on the third term in the RHS of (J.278) for $\alpha > 0$:

$$N_3(t) := \gamma^2 c^2 \int_0^t \tilde{\sigma}^2(s)\mathcal{J}(cF(t) - cF(s))\mathsf{d}s, \tag{J.288}$$

where we have

$$\rho^2\tilde{\sigma}^2(s) = 2\alpha\tilde{\eta}(0)^2(1-s)^{2\alpha-1}. \tag{J.289}$$

The change of variable $u = cF(t) - cF(s)$ yields

$$N_3(t) = \gamma^2 c \int_0^{cF(t)} \frac{\tilde{\sigma}^2(s(u))}{\tilde{\eta}(s(u))}\mathcal{J}(u)\mathsf{d}u \asymp \frac{\gamma^2}{\rho^2} v \int_0^{cF(t)} (1-s(u))^{\alpha-1}\mathcal{J}(u)\mathsf{d}u, \tag{J.290}$$

which gives, plugging in (J.283),

$$N_3(t) \asymp \frac{\gamma^2}{\rho^2} v \int_0^{cF(t)} \left(\frac{u}{v} + (1-t)^{\alpha+1} \right)^{\frac{\alpha-1}{\alpha+1}} \mathcal{J}(u) \mathsf{d}u. \tag{J.291}$$

Using Proposition J.20, we have

$$\begin{aligned} N_3(1) &\asymp \frac{\gamma^2}{\rho^2} v^{\frac{2}{\alpha+1}} \left(\int_0^1 u^{\frac{\alpha-1}{\alpha+1}} \mathcal{J}(u) \mathsf{d}u + \int_1^{cF(1)} u^{\frac{\alpha-1}{\alpha+1}} \mathcal{J}(u) \mathsf{d}u \right) \\ &\asymp \frac{\gamma^2}{\rho^2} v^{\frac{2}{\alpha+1}} \left(\int_0^1 u^{\frac{\alpha-1}{\alpha+1}} \mathsf{d}u + \int_1^{cF(1)} u^{\frac{\alpha-1}{\alpha+1}} u^{-\kappa_3} \mathsf{d}u \right). \end{aligned} \tag{J.292}$$

Notice that, since $cF(1) \asymp v = \omega_\gamma(1)$ and $\kappa_3 = 2 - \phi$, we have

$$\int_1^{cF(1)} u^{\frac{\alpha-1}{\alpha+1}} u^{-\kappa_3} \mathsf{d}u \asymp \begin{cases} 1 & \text{if } \dfrac{\alpha-1}{\alpha+1} + \phi < 1, \\ v^{\frac{\alpha-1}{\alpha+1}+\phi-1} & \text{if } \dfrac{\alpha-1}{\alpha+1} + \phi > 1, \\ \ln v & \text{if } \dfrac{\alpha-1}{\alpha+1} + \phi = 1. \end{cases} \tag{J.293}$$

This yields, since $\int_0^1 u^{\frac{\alpha-1}{\alpha+1}} \mathsf{d}u \asymp 1$ for $\alpha > 0$,

$$N_3(1) \asymp \begin{cases} \dfrac{\gamma^2}{\rho^2} v^{\frac{2}{\alpha+1}} & \text{if } \dfrac{\alpha-1}{\alpha+1} + \phi < 1, \\ \dfrac{\gamma^2}{\rho^2} v^{\phi} & \text{if } \dfrac{\alpha-1}{\alpha+1} + \phi > 1, \\ \dfrac{\gamma^2}{\rho^2} v^{\phi} \ln v & \text{if } \dfrac{\alpha-1}{\alpha+1} + \phi = 1. \end{cases} \tag{J.294}$$

Thus, if $\alpha \geq 1$, it is sufficient to have $\gamma^2 v/\rho^2 \lesssim 1$ to guarantee that $N_3(1) \lesssim 1$ (recall that $\phi < 1$). Notice that, a direct computation from (J.291) shows that this condition is sufficient to guarantee $N_3(t) \lesssim 1$ uniformly in $t \in [0,1]$.

In the case $\alpha = 1/2$, only the first case in (J.294) is possible, and this gives the condition $\gamma^2 v^{4/3}/\rho^2 \lesssim 1$ to guarantee that $N_3(1) \lesssim 1$ (and this is also sufficient to guarantee $N_3(t) \lesssim 1$ uniformly in $t \in [0,1]$).

Thus, the first part of the thesis follows from plugging (J.280), (J.286) and (J.294) in (J.278), since for $\alpha = 0$ we have $v = c\tilde{\eta}(1)$, and $\tilde{\eta}(1) = 0$ for $\alpha > 0$.

The second part of the thesis follows from the fact that the RHS of (J.278) is uniformly $\lesssim 1$ due to (J.280) and (J.287) for *(i)* $\alpha = 0$, or *(ii)* $\alpha = 1/2$ and $\gamma^2 v^{4/3}/\rho^2 \lesssim 1$, or *(iii)* $\alpha \geq 1$ and $\gamma^2 v/\rho^2 \lesssim 1$, due to the discussion following (J.294). □

Lemma J.24. *Let $\tilde{\eta}(t) = \tilde{\eta}(0)(1-t)^\alpha$, with $\alpha = \{0, 1/2\}$ and $\alpha \geq 1$, such that $\tilde{\eta}(0) \leq 2/\gamma$, and denote $v = \tilde{\eta}(0)c$. Assume $\rho \asymp \gamma^b$ with $b < 1$ independent of γ.*

Suppose $\phi(\alpha+1) < 2$. Set $c \lesssim 1$ and $c\tilde{\eta}(0) = \gamma^a$, with

$$a = \begin{cases} -\dfrac{\alpha+1}{K+1} & \text{if } b \leq \dfrac{K}{2(K+1)}, \\ -\dfrac{2(1-b)(\alpha+1)}{K+2} & \text{if } b > \dfrac{K}{2(K+1)}, \end{cases} \tag{J.295}$$

where $K = (2-\phi-\psi)(\alpha+1)$. *Then, we have that* $R(1) + 2c^2\tilde{\eta}^2(1)\gamma^2/\rho^2 \asymp \gamma^h$, *with*

$$h = \begin{cases} \dfrac{K}{K+1} & \text{if } b \leq \dfrac{K}{2(K+1)}, \\ \dfrac{2K(1-b)}{K+2} & \text{if } b > \dfrac{K}{2(K+1)}. \end{cases} \tag{J.296}$$

Suppose $\phi(\alpha+1) \geq 2$. *Set* $c \lesssim 1$, *and* $c\tilde{\eta}(0) = \gamma^a$. *If*

$$b \leq 1 - \frac{(\kappa_1+\phi)(\alpha+1)}{2(K+1)}, \tag{J.297}$$

then set a *as in the first case of* (J.295) *to achieve the same* h *as in the first case of* (J.296). *If* (J.297) *does not hold, then set*

$$a = -\frac{2(1-b)}{\kappa_1+\phi}, \tag{J.298}$$

which gives $R(1) + 2c^2\tilde{\eta}^2(1)\gamma^2/\rho^2 \asymp \gamma^h$, *with*

$$h = \begin{cases} \dfrac{2\kappa_1(1-b)}{\kappa_1+\phi} & \text{if } \phi(\alpha+1) > 2, \\ \dfrac{2\kappa_1(1-b)}{\kappa_1+\phi} + \dfrac{\ln(a\ln\gamma)}{\ln\gamma} & \text{if } \phi(\alpha+1) = 2. \end{cases} \tag{J.299}$$

Furthermore, for each value of α, *we have* $R(1) + 2c^2\tilde{\eta}^2(1)\gamma^2/\rho^2 \gtrsim \gamma^h$ *for all choices of* c *and all choices of* a *independent of* γ.

Proof. Consider the case $c = \omega_\gamma(1)$. Due to Lemma J.5 and the same argument used in (J.142), we have that

$$\begin{aligned} R(t) &\gtrsim \mathcal{F}(\Gamma(t)) + \int_0^t \Big(\tilde{\eta}^2(s)\gamma\mathcal{K}(\Gamma(t)-\Gamma(s)) + c^2\tilde{\sigma}^2(s)\gamma^2\mathcal{J}(\Gamma(t)-\Gamma(s))\Big)\,\mathrm{d}s \\ &\gtrsim \mathcal{F}(F(t)) + \int_0^t \Big(\tilde{\eta}^2(s)\gamma\mathcal{K}(F(t)-F(s)) + c^2\tilde{\sigma}^2(s)\gamma^2\mathcal{J}(F(t)-F(s))\Big)\,\mathrm{d}s \\ &\gtrsim \mathcal{F}(F(t)) + \int_0^t \Big(\tilde{\eta}^2(s)\gamma\mathcal{K}(F(t)-F(s)) + \tilde{\sigma}^2(s)\gamma^2\mathcal{J}(F(t)-F(s))\Big)\,\mathrm{d}s, \end{aligned} \tag{J.300}$$

where the the second step holds for an argument similar to the one in the proof of Lemma J.21, as $\mu_c(R) < 1$. Then, the risk at all times is lower bounded by an expression which is identical in form to the setting $c \asymp 1$, which implies that the smallest rates for $R(1) + 2c^2\tilde{\eta}^2(1)\gamma^2/\rho^2$ can be achieved for $c \lesssim 1$. Furthermore, if $c \lesssim 1$ and $v \lesssim 1$, (J.280) gives $R(1) \gtrsim 1$, which is a vacuous rate.

Then, let us focus on $c \lesssim 1$ and $v = \omega_\gamma(1)$. Then, Lemma J.23 gives the lower bound in (J.276). During the proof, we will suppose this bound to be tight, and we will find the value of a (with $v = \gamma^a$) that minimizes this lower bound. Later, we will verify that $a \in [-1, 0)$ (such that it is possible to consider $c \lesssim 1$ and $\tilde{\eta}(0) < 2/\gamma$), $2 - 2b + 4a/3 \geq 0$ for $\alpha = 1/2$, and $2 - 2b + a \geq 0$ for $\alpha \geq 1$, guaranteeing that indeed the lower bound in (J.276) is tight and that the choice of a is optimal.

- Suppose $\phi(\alpha+1)<2$. Then, Lemma J.23 gives

$$R(1)+\frac{2c^2\tilde{\eta}^2(1)\gamma^2}{\rho^2}\gtrsim\gamma^{e_1}+\gamma^{e_2}+\gamma^{e_3}, \tag{J.301}$$

with

$$e_1=-a\kappa_1,\qquad e_2=\frac{a}{\alpha+1}+1,\qquad e_3=2-2b+\frac{2a}{\alpha+1}. \tag{J.302}$$

Then, the problem of hyper-parameter optimization reduces to finding a such that $\min(e_1,e_2,e_3)$ is maximized, giving the minimum lower bound on the risk. Consider the value of a for which $e_1=e_2$. This gives

$$a=-\frac{\alpha+1}{K+1}\in[-1,0], \tag{J.303}$$

where the inclusion follows from $\alpha\geq 0$ and $K>\alpha$ since $\kappa_1>1$. Note that, for this value of a, if $b\leq\frac{K}{2(K+1)}$, then we have

$$e_1=e_2=\frac{K}{K+1}\leq e_3. \tag{J.304}$$

Then, since e_1 is monotonically decreasing in a and e_2 and e_3 are increasing in a (see Figures J.2, J.4 and J.5 for a graphical presentation of this case), (J.303) gives the optimal choice of a and (J.304) the corresponding lower bound on the risk. Furthermore, note that, in this case, we have

$$2-2b\geq\frac{K+2}{K+1}=1+\frac{1}{K+1}=1-\frac{a}{\alpha+1}. \tag{J.305}$$

Then, for $\alpha\geq 1$ this gives $2-2b+a\geq 0$, and for $\alpha=1/2$ this gives $2-2b+4a/3\geq 0$, which implies that (J.301) is tight, due to the second part of the thesis of Lemma J.23.

If $b>\frac{K}{2(K+1)}$, the value of a in (J.303) is such that $e_1=e_2\geq e_3$. Then, set $e_1=e_3$, which yields

$$a=-\frac{2(1-b)(\alpha+1)}{K+2}\in[-1,0], \tag{J.306}$$

where the inclusion follows from $b\leq 1$ and $\kappa_1>1$. Note that, for this value of a, since $b>\frac{K}{2(K+1)}$ we have

$$e_1=e_3=\frac{2K(1-b)}{K+2}\leq e_2. \tag{J.307}$$

Then, since e_1 is monotonically decreasing in a and e_2 and e_3 are increasing in a (see Figure J.3 for a graphical presentation of this case), (J.306) gives the optimal choice of a and (J.307) the corresponding lower bound on the risk. Furthermore, in this case, we have that

$$2-2b+a=\frac{2(1-b)(K+1-\alpha)}{K+2}>0, \tag{J.308}$$

where the last step holds since $K>\alpha$. If $\alpha=1/2$, we have that $2-2b+4a/3>0$ since $3\kappa_1/2+2>3/2$, which implies that (J.301) is tight due to the second part of the thesis of Lemma J.23. This concludes the argument for the case $\phi(\alpha+1)<2$.

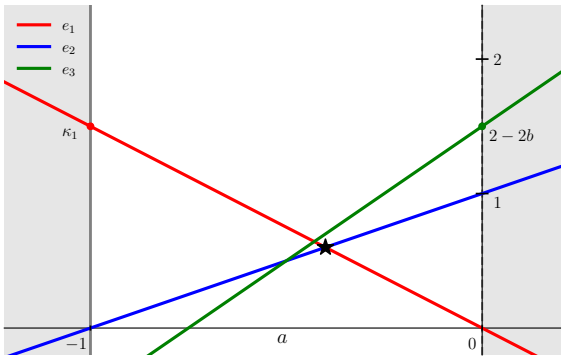


Figure J.2: Plots of e_1, e_2 and e_3 as in (J.302) as a function of a for $\kappa_1 = 1.5$, $\alpha = 0$, $b = 0.25$. Privacy comes for free, since $\arg\max(\min(e_1, e_2, e_3)) = \arg\max(\min(e_1, e_2))$.

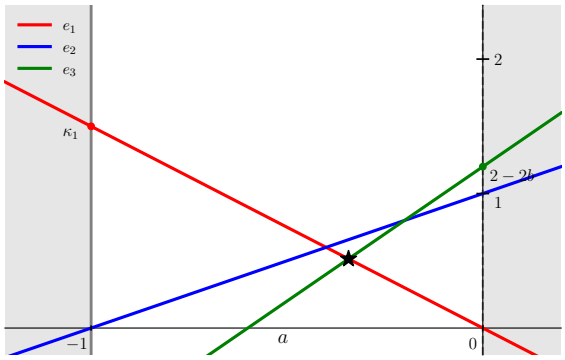


Figure J.3: Plots of e_1, e_2 and e_3 as in (J.302) as a function of a for $\kappa_1 = 1.5$, $\alpha = 0$, $b = 0.4$. Privacy comes with a performance cost, since $\arg\max(\min(e_1, e_2, e_3)) = \arg\max(\min(e_1, e_3))$.

- Suppose $\phi(\alpha+1) > 2$. Then, Lemma J.23 gives that (J.301) holds with

$$e_1 = -a\kappa_1, \qquad e_2 = \frac{a}{\alpha+1} + 1, \qquad e_3 = 2 - 2b + a\phi. \tag{J.309}$$

As before, set a for which $e_1 = e_2$. This gives the result in (J.303), and if

$$b \leq 1 - \frac{(\kappa_1 + \phi)(\alpha+1)}{2(K+1)}, \tag{J.310}$$

we have that (J.304) holds. Notice that the case $\phi(\alpha+1) > 2$ cannot happen for either $\phi \leq 0$ or $\alpha \leq 1$, and for $\alpha > 1$, we have

$$2 - 2b + a \geq \frac{(\kappa_1+\phi)(\alpha+1)}{K+1} + a = \frac{(\kappa_1+\phi-1)(\alpha+1)}{K+1} = \frac{(1-\psi)(\alpha+1)}{K+1} > 0, \tag{J.311}$$

where the last step holds since $1 - \psi > \phi$ by Assumption 9.13. Then, this implies that (J.301) is tight due to the second part of the thesis of Lemma J.23.

If (J.310) does not hold, the value of a in (J.303) is such that $e_1 = e_2 \geq e_3$. Then, set $e_1 = e_3$, which yields

$$a = -\frac{2(1-b)}{\kappa_1 + \phi} \in [-1, 0], \tag{J.312}$$

where the inclusion follows from $b \leq 1$ and the lower bound on b. Notice that, for this value of a and interval of b we have

$$e_1 = e_3 = \frac{2\kappa_1(1-b)}{\kappa_1+\phi} \leq e_2. \tag{J.313}$$

Then, since e_1 is monotonically decreasing in a and e_2 and e_3 are increasing in a, (J.312) gives the optimal choice of a and (J.313) the corresponding lower bound on the risk. Furthermore, in this case, we have that

$$2 - 2b + a = \frac{2(1-b)(\kappa_1+\phi-1)}{\kappa_1+\phi} > 0, \tag{J.314}$$

where the last step holds since $\phi > 0$, which implies that (J.301) is tight due to the second part of the thesis of Lemma J.23. This proves the thesis for the case $\phi(\alpha+1) > 2$.

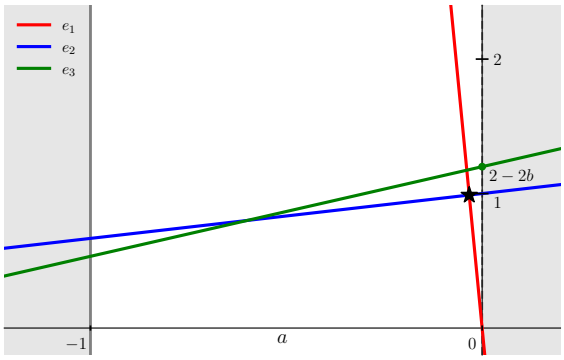


Figure J.4: Plots of e_1, e_2 and e_3 as in (J.302) as a function of a for $\kappa_1 = 30$, $\alpha = 2$, $b = 0.4$. Privacy comes for free, since $\arg\max(\min(e_1, e_2, e_3)) = \arg\max(\min(e_1, e_2))$.

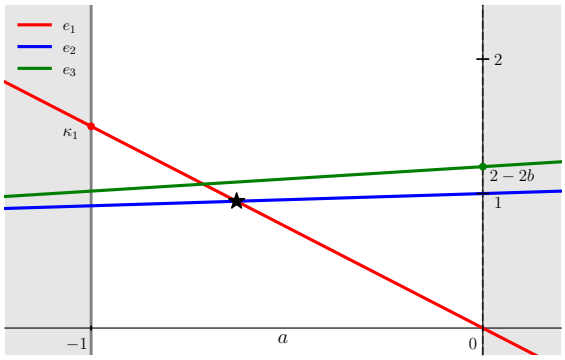


Figure J.5: Plots of e_1, e_2 and e_3 as in (J.302) as a function of a for $\kappa_1 = 1.5$, $\alpha = 10$, $b = 0.4$. Privacy comes for free, since $\arg\max(\min(e_1, e_2, e_3)) = \arg\max(\min(e_1, e_2))$.

- Suppose $\phi(\alpha+1) = 2$. Then, Lemma J.23 gives that (J.301) holds with

$$e_1 = -a\kappa_1, \qquad e_2 = \frac{a}{\alpha+1} + 1, \qquad e_3 = 2 - 2b + a\phi - \frac{\ln(|a|\ln(1/\gamma))}{\ln(1/\gamma)}. \tag{J.315}$$

  Since the last term on the RHS of the last expression is $o_\gamma(1)$, the case given by (J.310) is identical to the previous one.

  When instead (J.310) does not hold, setting a as in (J.312) is also optimal as $\gamma \to 0$, since $\ln(1/\gamma) = \omega_\gamma(\ln(|a|\ln(1/\gamma))$. Thus, the desired result follows.

□

Proof of Theorem 9.14. The proof is a direct consequence of Lemma J.24 together with Theorem 9.6 and Proposition 9.8, after replacing $\kappa_1 = 2 - \phi - \psi$. □

Proof of Corollary 9.15. If $\phi \leq 0$, for any $\alpha \geq 0$ we have that $\phi(\alpha+1) \leq 0 < 2$. Then, taking $\alpha \to \infty$ (and consequently $K \to \infty$ as defined in Lemma J.24), h as defined in (J.296) converges (from below) to $\min(1, 2(1-b))$, which gives the first part of the thesis.

If $\phi > 0$, it is still desirable to consider the limit $\alpha \to \infty$, as h is non-decreasing in α (this can be seen from Lemma J.23). Then, this setting reduces to the case $\phi(\alpha+1) > 2$. As $\alpha \to \infty$, we have that the threshold in (J.310) approaches

$$b^* = 1 - \frac{\kappa_1 + \phi}{2\kappa_1} = \frac{1}{2} - \frac{\phi}{2(2-\phi-\psi)} \tag{J.316}$$

from above. Furthermore, the exponent h approaches 1 from below if $b \leq b^*$, and it is equal to $2(1-b)^{\frac{2-\phi-\psi}{2-\psi}}$ otherwise, which concludes the proof. □