PNAS RESEARCH ARTICLE | APPLIED MATHEMATICS OPEN ACCESS

Privacy for free in the overparameterized regime

Simone Bombari[a,1] and Marco Mondelli[a,1]

Edited by David Weitz, Harvard University, Cambridge, MA; received November 7, 2024; accepted March 9, 2025

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 has led to upper bounds on the excess population risk $\mathcal{R}_P$ in different learning settings. However, such bounds typically degrade with overparameterization, 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 work, 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., $|\mathcal{R}_P| = o(1)$, 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 overparameterization inherently hinders performance in private learning.

differential privacy | deep learning | overparameterization | differentially private gradient descent | random features model

Significance

In many deep learning applications, training datasets routinely include personal, sensitive information. Learning from these data is possible without creating privacy infringement via methods guaranteeing differential privacy, designed to provide provable protection to any individual user. However, differential privacy comes with a performance cost, and the cost is often believed to grow with the number of parameters of the learning model. Our work challenges this view, showing that overparameterization is not at odds with privacy. In fact, we prove that, for a class of overparameterized models having access to enough training samples, privacy even comes for free, i.e., with a small loss in performance. This result provides theoretical support to the development of differentially private models at scale.

Deep learning models are vulnerable to attacks directed to retrieve information about the training dataset (1, 2), which is concerning when sensitive data are included in the learning pipeline. To allow the usage of such data, differential privacy (DP) (3) consolidated as the golden standard for privacy. This framework comes with algorithms (4) that provide formal protection guarantees for each sample in the training set, which is safeguarded (up to some level) by any adversary with access to the trained model and the rest of the dataset. Specifically, neural networks are trained in a differentially private way via e.g. DP (stochastic) gradient descent (DP-GD) (4). This 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*. However, privacy guarantees often come with a performance cost with respect to standard GD (5). Furthermore, private training involves carefully tuning additional hyperparameters, e.g., clipping constant, noise magnitude, and number of training iterations, which increases the computational cost, also due to the higher training times and memory loads of DP optimization (6).

The challenging problem of optimizing neural networks with an assigned privacy guarantee has motivated a thriving field of research proposing architectures and training algorithms (5, 7, 8). Concurrently, theoretical studies have emerged with the scope of quantifying privacy–utility tradeoffs. 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* 2.1. To provide meaningful protection, practitioners pick constant-order values of ε ($\varepsilon \in \{1, 2, 4, 8\}$) and $\delta < 1/n$, where n is the number of training samples (9). Utility is typically measured as the degradation in generalization of the DP solution $\theta^p \in \mathbf{R}^p$ compared to a nonprivate baseline $\theta^* \in \mathbf{R}^p$, where p is the number of parameters of the model. Considering the standard supervised setting and denoting by $(x, y) \sim \mathcal{P}_{XY}$ an input–label pair with distribution $\mathcal{P}_{XY}$, the excess population risk is defined as

$$\mathcal{R}_P = \mathbb{E}_{(x,y)\sim\mathcal{P}_{XY}}\left[\ell(x, y, \theta^p)\right] - \mathbb{E}_{(x,y)\sim\mathcal{P}_{XY}}\left[\ell(x, y, \theta^*)\right], \quad [1]$$

where $\ell(x, y, \theta)$ is the loss over the sample (x, y) of the model evaluated in θ. Intuitively, $\mathcal{R}_P$ worsens with more stringent privacy requirements on θ^p (i.e., smaller values of ε, δ), and a rich line of work spanning over a decade has investigated the trade-off (9–15). Despite this flurry of research, existing results are unable to address the overparameterized regime, i.e., $p = \Omega(n)$, as bounds on $\mathcal{R}_P$ become vacuous (see the comparison with

Author affiliations: [a]Institute of Science and Technology Austria, Klosterneuburg 3400, Austria

Author contributions: S.B. and M.M. designed research; performed research; analyzed data; and wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: simone.bombari@ista.ac.at or marco.mondelli@ista.ac.at.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2423072122/-/DCSupplemental.

Published April 11, 2025.

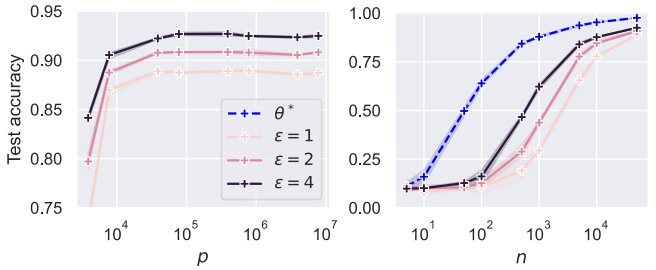

Fig. 1. Test accuracy of DP-GD on MNIST for a 2-layer, fully connected ReLU network, plotted as a function of (*Left*) the number of parameters p with fixed $n = 50{,}000$, and (*Right*) the number of training samples n with fixed hidden layer width $= 1{,}000$. Further details on the experimental setting can be found in Section 3.

previous work below). This is sometimes understood via the qualitative argument that the noise introduced by DP-GD increases with the dimension of the parameter space (16, 17), and it has lead to DP algorithms acting on lower dimensional subspaces (7, 8, 18, 19).

On the other hand, empirical evidence that larger models are beneficial on down-stream tasks requiring private fine-tuning is provided in refs. 20 and 21, which motivated theoretical studies giving refined privacy–utility tradeoffs (22). Perhaps surprisingly, the recent work (6) gives evidence of the benefits of scale even in the absence of public pretraining data, as the generalization performance improves with model size on CIFAR-10 and ImageNet, following an accurate hyperparameter search. In the *Left* panel of Fig. 1, we investigate the interplay between privacy and overparameterization in a simpler and more controllable setting: training a 2-layer, fully connected ReLU network on MNIST with DP-GD (*Algorithm* 1). We vary the network width, spanning both the underparameterized and overparameterized regime, questioning whether the algorithm suffers as the number of parameters grows. The plot shows that this is not the case: The test accuracy increases until the network is wide enough and then plateaus. Furthermore, the gap between the GD solution θ^* and the DP-GD one tends to vanish by increasing the number of training samples n; see the *Right* panel of Fig. 1.

For nonprivate optimization, the apparent contradiction between the excellent generalization of overparameterized models and the classical bias–variance tradeoff has been the subject of intense investigation, highlighting e.g. the role of benign overfitting (23, 24) and double descent (25, 26). The phenomenology discussed above is hard to explain given the current theoretical understanding of overparameterized private training. Thus, this calls for a framework able to i) provide generalization guarantees, and ii) characterize how the hyperparameters of DP-GD affect performance.

1.1. Informal Overview. In this work, we provide privacy–utility guarantees $\mathcal{R}_P = o(1)$ under overparameterization—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 overparameterized regime.

We consider a family of models where the number of parameters p can be significantly larger than the number of samples n and the input dimension d, as in Fig. 1. Specifically, we focus on the widely studied random features (RF) model (27) with quadratic loss, which takes the form

$$\ell(x, y, \theta) = (f_{\mathrm{RF}}(x, \theta) - y)^2, \qquad f_{\mathrm{RF}}(x, \theta) = \phi(Vx)^\top \theta, \quad \textbf{[2]}$$

where f_{RF} is a generalized linear model, $\phi : \mathbf{R} \to \mathbf{R}$ a nonlinearity applied component-wise to the vector $Vx \in \mathbf{R}^p$, and $V \in \mathbf{R}^{p \times d}$ a random weight matrix. The RF model can be regarded as a 2-layer network, where only the second layer θ is trained and V is 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 (23, 26). The DP-trained solution θ^p is obtained via DP-GD (*Algorithm* 1), and the nonprivate baseline θ^* in Eq. **1** via (nonprivate) GD. At this point, we can present an informal version of our result (formally stated in Section 2).

Theorem 1 (informal). *Consider the RF model in Eq. **2** with input dimension d and number of parameters p. Let n be the number of training samples and $\mathcal{R}_P$ be defined according to Eq. **1**, where θ^* is the solution of GD and θ^p is the (ε, δ)-differentially private solution of DP-GD (Algorithm 1). Then, for all sufficiently overparameterized models, under some technical conditions, the following holds with high probability*

$$|\mathcal{R}_P| = \tilde{\mathcal{O}}\left(\frac{d}{n\varepsilon} + \sqrt{\frac{d}{n}} + \sqrt{\frac{n}{d^{3/2}}}\right) = o(1). \qquad \textbf{[3]}$$

In words, in the regime $d \ll n \ll d^{3/2}$—considered e.g. in refs. 28 and 29; see Eq. **6** for details—we show that $|\mathcal{R}_P| = o(1)$ 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 Eq. **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 Eq. **3** does not depend on p as we only require a lower bound on it (Eq. **6**). The dependence of $\mathcal{R}_P$ on δ is only logarithmic and it is neglected in the notation $\tilde{\mathcal{O}}(\cdot)$ that hides polylogarithmic factors in δ and n.

1.2. Comparison with Previous Work. In the setting of convex unconstrained optimization, (12) focuses on generalized linear models (GLMs, i.e., $\ell(x, y, \theta)$ is replaced by $\ell(\varphi(x)^\top \theta, y)$, where φ is the feature map), assumes L-Lipschitz loss with strongly convex regularization, and bounds the excess population risk with respect to the Bayes optimal solution θ^* as $\mathcal{R}_P = \tilde{\mathcal{O}}(L\|\theta^*\|_2/\sqrt{n\varepsilon})$. (14) lifts the assumption on the strong convexity and improves the previous bound via the projector M on the column space of $\mathbb{E}_{x\sim\mathcal{P}_X}[\varphi(x)\varphi(x)^\top]$. (22) considers Lipschitz losses with ℓ_2 regularization and recovers the bound of ref. 14 for GLMs. A similar approach is taken by Ma et al. (30) that removes the dependence on p at the cost of an additional factor $\mathrm{tr}(\tilde{H})$, where $\tilde{H} \succeq \sup_\theta \nabla^2_\theta \mathbb{E}_{(x,y)\sim\mathcal{P}_{XY}}[\ell(x, y, \theta)]$. (31) considers GLMs and it also recovers the result of ref. 14 in the Lipschitz setting with $\varepsilon = \Omega(1)$.

Importantly, even for an RF model, existing bounds do not access the overparameterized regime: Taking ε of constant order makes the upper bounds on the excess population risk to read (at best) $\mathcal{R}_P = \mathcal{O}(1)$, which is vacuous as the performance of a trivial model outputting zero is of the same order. We now explain why this is the case. First, note that the RF model in Eq. **2** has a non-Lipschitz (quadratic) loss, which does not allow a direct application of most previous bounds. 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 $\|\theta^*\|_2$. This provides the scaling of an effective Lipschitz constant of the

Algorithm 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**

$$g(x_i, y_i, \theta_{t-1}) \leftarrow \nabla_\theta \ell(x_i, y_i, \theta_{t-1})$$

$$g_{C_{\text{clip}}}(x_i, y_i, \theta_{t-1}) \leftarrow \frac{g(x_i, y_i, \theta_{t-1})}{\max\left(1, \frac{\|g(x_i, y_i, \theta_{t-1})\|_2}{C_{\text{clip}}}\right)}$$

$$g_{\theta_{t-1}} \leftarrow \frac{1}{n}\sum_i g_{C_{\text{clip}}}(x_i, y_i, \theta_{t-1})$$

$$\theta_t = \theta_{t-1} - \eta g_{\theta_{t-1}} + \sqrt{\eta}\,\frac{2C_{\text{clip}}}{n}\sigma\mathcal{N}(0, I_p)$$

Output: Model parameters θ_T

model, as if the optimization was bounded to the set $\mathcal{B}$. From our results in later sections, it can be shown that $\|\theta^*\|_2 = \Theta(\sqrt{n/p})$, which gives $\|\theta^*\|_2 \sup_{\theta\in\mathcal{B}} \|\nabla_\theta \ell(x_i, y_i, \theta)\|_2 = \Theta(n)$. This trivializes the bound in ref. 12 to $\mathcal{R}_{\text{P}} = \tilde{\mathcal{O}}(\sqrt{n}/\varepsilon)$. Furthermore, even by assuming that the loss function in Eq. **2** is Lipschitz, the result improves only by a factor $\sqrt{n}$, i.e., $\mathcal{R}_{\text{P}} = \tilde{\mathcal{O}}(1/\varepsilon)$, which is again trivial when $\varepsilon = \Theta(1)$. Similar considerations apply to the bounds in refs. 14, 22, and 31. As concerns the result in ref. 30, it can be verified that $\text{tr}(\tilde{H}) \geq \mathbb{E}_x\left[\|\varphi(x)\|_2^2\right] = \Theta(p)$, which reintroduces the dependence on the number of parameters p, preventing an improvement upon (12). Finally, while (31) gives 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 *SI Appendix*, section 1.A, together with an additional review of the related literature [e.g., on constrained optimization (9, 13), parameter estimation (32, 33), and linear regression (15, 34)].

2. Setting and Main Result

2.1. Differential Privacy (DP) and DP-GD. The definition of DP builds on the notion of adjacent datasets. In our setting, a dataset D' is said to be adjacent to a dataset D if they differ by only one sample.

Definition 2.1 [(ε, δ)-DP (3)]. A randomized algorithm $\mathcal{A} : \mathcal{D} \to \mathbf{R}^p$ satisfies (ε, δ)-differential privacy if for any pair of adjacent datasets $D, D' \in \mathcal{D}$ and for any $S \subseteq \mathbf{R}^p$, we have

$$\mathbb{P}\left(\mathcal{A}(D) \in S\right) \leq e^{\varepsilon}\mathbb{P}\left(\mathcal{A}(D') \in S\right) + \delta. \qquad [4]$$

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'.

A popular choice to enforce (ε, δ)-DP relies on DP-GD algorithms, which perturb individual updates during training, providing privacy guarantees based on the size of the perturbation and the number of iterations (4, 13, 35). In this work, we consider *Algorithm* 1, a variant of the well-established method in ref. 4 without stochastic batching (not considered for simplicity). Its privacy guarantees are below, with the complete argument deferred to *SI Appendix*, section 2.

Proposition 2.2. *For any* $\delta \in (0, 1)$, $\varepsilon \in (0, 8\log(1/\delta))$, *if we set*

$$\sigma \geq \sqrt{\eta T}\frac{\sqrt{8\log(1/\delta)}}{\varepsilon}, \qquad [5]$$

then Algorithm 1 is (ε, δ)*-differentially private.*

2.2. Problem Setup. We consider *Algorithm* 1 in an overparameterized RF model. For simplicity, we set the initialization $\theta_0 = 0$. The random features matrix $V \in \mathbf{R}^{p\times d}$ is i.i.d. Gaussian, i.e., $V_{ij} \sim_{\text{i.i.d.}} \mathcal{N}(0, 1/d)$. We assume the n training samples $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ to be i.i.d. taken from the joint distribution $\mathcal{P}_{XY}$, such that the labels $(y_1, \ldots, y_n)$ are bounded and the marginal $\mathcal{P}_X$ satisfies the following properties: i) $x \sim \mathcal{P}_X$ is sub-Gaussian, with $\|x\| = \mathcal{O}(1)$; ii) the data $x \sim \mathcal{P}_X$ are normalized, i.e., $\|x\|_2 = \sqrt{d}$; iii) $\lambda_{\min}\left(\mathbb{E}_{x\sim\mathcal{P}_X}\left[xx^\top\right]\right) = \Omega(1)$, i.e., the second-moment matrix is well conditioned. Taken together, these requirements are implied by the stronger conditions in refs. 26 and 28. Furthermore, they are fulfilled by normalized multivariate Gaussians with well-conditioned covariance and by the normalized features of a class of fully connected neural networks (36). We note that the adjacency of *Definition* 2.1 is not subject to our distributional assumptions and *Proposition* 2.2 holds with D and D' differing by a sample in any arbitrary way.

The scaling of input data and random features V guarantees that the preactivations of the model (i.e., the entries of Vx) are of constant order. We then process the entries of Vx via an activation function $\phi : \mathbf{R} \to \mathbf{R}$, which we require to be nonlinear, Lipschitz continuous, and such that $\mu_0 = \mu_2 = 0$, and $\mu_1 \neq 0$, where μ_k denotes its k-th Hermite coefficient. 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 ref. 28, at the cost of a more involved analysis.

We further consider the following scaling of the problem

$$n = \mathcal{O}\left(\sqrt{p}\right), \quad n = \omega\left(d\log^2 d\right), \quad n = o\left(\frac{d^{3/2}}{\log^3 d}\right). \qquad [6]$$

To guarantee that the RF model interpolates the data, it suffices that $p \gg n$ (see e.g. refs. 28 and 37), and we expect our result to hold under this milder assumption on p as well. We also assume $\log n = \Theta(\log p)$, which is mild and could be relaxed at the expenses of a polylogarithmic dependence on p in our final result in *Theorem* 2. We finally remark that the regime $d \ll n \ll d^{3/2}$ corresponds to standard datasets, such as CIFAR-10 ($n = 5\cdot 10^4$, $d \approx 3\cdot 10^3$), or ImageNet as considered in ref. 38 ($n \approx 1.3\cdot 10^6$, $d \approx 9\cdot 10^4$).

According to *Proposition* 2.2, we consider values of the privacy budget $\delta \in (0, 1)$, $\varepsilon \in (0, 8\log(1/\delta))$, and

$$\frac{\varepsilon}{\sqrt{\log(1/\delta)}} = \omega\left(\frac{d\log^5 n}{n}\right). \qquad [7]$$

This lower bound on ε still allows for strong privacy regimes with $\varepsilon = o(1)$, as $n \gg d$ from Eq. **6**. We set the hyperparameters of *Algorithm* 1 as

$$T = \frac{d\log^2 n}{\eta p}, \qquad C_{\text{clip}} = \sqrt{p}\log^2 n, \qquad [8]$$

with σ given by the RHS in *Proposition* 2.2, which guarantees that *Algorithm* 1 is (ε, δ)-DP. We also define the nonprivate baseline as the solution of the gradient flow equation

$$\mathrm{d}\hat{\theta}(t) = -\nabla\mathcal{L}(\hat{\theta}(t))\mathrm{d}t, \qquad \theta^* = \lim_{t\to+\infty}\hat{\theta}(t), \qquad [9]$$

where $\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\ell(\varphi(x_i)^\top\theta - y_i)$ with $\varphi(x_i) := \phi(Vx_i)$ is the training loss (see section 5.1 of ref. 23 for details on the convergence). Then, our main result is formally stated below.

Theorem 2. *Consider the RF model in Eq. **2** with input dimension d and number of features p. Let n be the number of training samples and $\mathcal{R}_P$ be defined in Eq. **1**, where θ^p is the solution θ_T of Algorithm 1, and θ^* is defined in Eq. **9**. Then, as η goes to 0, we have that*

$$|\mathcal{R}_P| = \mathcal{O}\left(\frac{d}{n\epsilon}\log^5 n\sqrt{\log(1/\delta)} + \sqrt{\frac{d}{n}} + \sqrt{\frac{n\log^3 d}{d^{3/2}}}\right),$$

with probability at least $1 - 2\exp\left(-c\log^2 n\right)$, where c is an absolute constant.

Existing work studies $\mathcal{R}_P$ by i) bounding the excess empirical risk of θ^p via convex optimization techniques, and by ii) extending the result to the excess population risk via stability arguments (9, 14, 22). In contrast, we consider the continuous process defined by *Algorithm* 1 as $\eta \to 0$, for the RF model [Proposition 2.7 of *SI Appendix*, section 2 ensures that the limit is (ϵ, δ)-DP]. This allows the use of probabilistic tools that provide a refined control on the trajectory of DP-GD. This approach has proven useful in the non private setting (28, 39), and in this work we apply it to DP learning. We also note that obtaining *Theorem* 2 still required overcoming significant technical barriers: While Mei et al. and Hu et al. (28, 39) establish the asymptotic test error of θ^* at convergence, we need a nonasymptotic 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, as discussed in Section 4.

3. Numerical Results and Discussion

3.1. Overparameterization Not at Odds with Privacy. *Theorem* 2 proves that, in the RF model, overparameterization is not inherently detrimental to private learning. The first panel of Fig. 2 verifies this by plotting the test loss of an RF model trained on a synthetic dataset via DP-GD, as the number of parameters p increases. We also report the performance achieved by (nonprivate) GD, which provides the baseline θ^*. While the test loss of θ^* displays the typical double-descent curve (25, 26, 40), 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 loss of θ^*. This is in agreement with *Theorem* 2, which predicts a small performance gap between θ^p and θ^* for overparameterized models. 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 ref. 41 for a connection between ridge and early stopping and also the discussion after Lemma 3.4 of *SI Appendix*, section 3.

3.2. Role of Hyperparameters. Both in Fig. 1 and in the first panel of Fig. 2, the hyperparameters in DP-GD are chosen to maximize the validation performance. In the second panel of Fig. 2, we take $C_{\text{clip}} = 0.5\sqrt{p}$ and report the test loss as a function of the number of iterations T, setting the noise according to *Proposition* 2.2 to guarantee the desired privacy budget. As suggested by Eq. **8**, the optimal T minimizing the test error decreases with p. More specifically, as we rescale the plot putting $\eta Tp/d$ on the x-axis of the third panel, the curves collapse onto each other, confirming the accuracy of the proposed scaling. The heat-map in the fourth panel displays the results of a full hyperparameter grid search over (C_{clip}, T) for a fixed p and ϵ. We distinguish 4 regions in hyperparameters space: In the 1) *Top-Right*, we have very low utility due to the large noise, 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, 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. At the center of the panel, we have $C_{\text{clip}} \sim \sqrt{p}$ and $\eta T \sim d/p$ as in Eq. **8**, which lead to low test loss. Moving toward the 4) *Top-Left* there is no decrease in performance, which is in line with earlier empirical work (6, 20) 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, as shown in the rightmost panel. The implementation

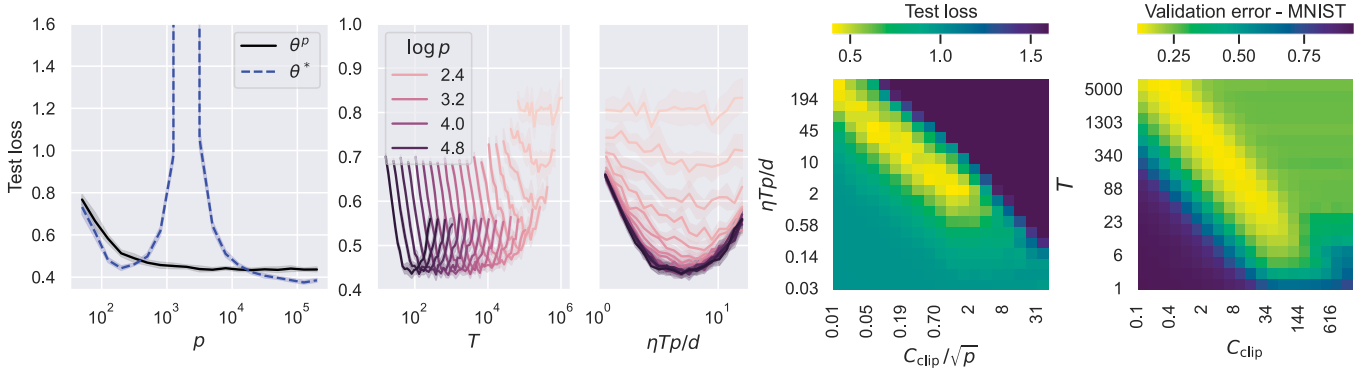

Fig. 2. Experiments on RF models with tanh activation and synthetic data sampled from a standard Gaussian distribution with $d = 100$ (first four panels), and on a 2-layer fully connected ReLU network trained with cross-entropy loss on MNIST ($d = 768$, $n = 50{,}000$) with privacy budget $\epsilon = 1$, $\delta = 1/n$ (last panel). For the RF model, the learning task is given by $y = \text{sign}(u^\top x)$, where $u \in \mathbf{R}^d$ is a fixed vector sampled from the unit sphere, and we consider a fixed number of training samples $n = 2{,}000$; θ^p is the solution of *Algorithm* 1 with $\varepsilon = 4$, $\delta = 1/n$, and θ^* is the solution of GD, both with small enough learning rate η. *First panel*: Test loss of θ^p and θ^* for different number of parameters p. *Second panel*: Test loss of θ^p as a function of the number of training iterations T. *Third panel*: Same plot as in the second panel, with x-axis set to $\eta Tp/d$. *Fourth panel*: Test loss of θ^p for a fixed $p = 40{,}000$, as a function of the hyperparameters (C_{clip}, T), in dark blue all values of the loss above 1.6. *Fifth panel*: Validation error for a fixed hidden-layer width set to 1,000 ($p \sim 10^6$), as a function of the hyperparameters (C_{clip}, T).

of the experiments is publicly available at the GitHub repository https://github.com/simone-bombari/privacy-for-free.

3.3. Privacy for Free. *Theorem* 2 proves that we can achieve privacy for free, which may seem surprising. The intuition is that in the RF model, when $n \gg d$, there is a surplus of samples that can be used to learn privately. In fact, the test error of (nonprivate) GD plateaus when n is between d and $d^{3/2}$ (28, 39), hence $\Theta(d)$ samples are enough to achieve utility, with the remaining ones leading to privacy. The *Right* panel of Fig. 1 displays the phenomenon. On the *Left*, the performance of θ^* increases with n, while private algorithms have low utility. Moving toward the right, the performance of GD saturates, and the utility of DP-GD increases, approaching the nonprivate baseline. The plateau in the test loss of GD has been shown for kernel ridge regression (28, 39, 42) and the RF model exhibits it 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 ref. 39. 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 the present paper, 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 3.5 in *SI Appendix*, section 3.

4. Methods

4.1. A Continuous View on DP-GD. Two challenges arise in the analysis of DP-GD as $\eta \to 0$: i) gradient clipping, and ii) noise injection. To overcome the former, we define a clipped loss $\mathcal{L}_{C_{\text{clip}}}(\theta)$. 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 $\sqrt{\eta}$ of the SD of the noise, as done e.g. in ref. 43.

As in ref. 14, we note that clipping in *Algorithm* 1 can be reformulated as optimizing the surrogate loss $\ell_{i,C_{\text{clip}}}(\varphi(x_i)^\top\theta - y_i)$, whose derivative for the i-th training sample reads

$$\ell'_{i,C_{\text{clip}}}(z) = \ell'(z)\min\left(1, \frac{C_{\text{clip}}}{|\ell'(z)|\,\|\varphi(x_i)\|_2}\right). \quad [10]$$

In other words, running *Algorithm* 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* 1 as

$$\theta_t - \theta_{t-1} = -\eta\nabla\mathcal{L}_{C_{\text{clip}}}(\theta_{t-1}) + \sqrt{\eta}\frac{2C_{\text{clip}}}{n}\sigma\mathcal{N}(0, I_p). \quad [11]$$

This update rule corresponds to the Euler–Maruyama discretization scheme of the SDE

$$\mathrm{d}\Theta(t) = -\nabla\mathcal{L}_{C_{\text{clip}}}(\Theta(t))\mathrm{d}t + \Sigma\,\mathrm{d}B(t), \quad [12]$$

with discretization η (see section 10.2 of ref. 44). Here, $B(t)$ is a p-dimensional Wiener process, $\Sigma := 2C_{\text{clip}}\sigma/n$, and the initial condition of Eq. **12** corresponds to the initialization of *Algorithm* 1, i.e., $\Theta(0) = \theta_0$. The strong convergence of the Euler–Maruyama method guarantees that, for any $\tau = \eta T$, θ_T from *Algorithm* 1 approaches $\Theta(\tau)$ from Eq. **12** as η gets smaller. We note that previous work (45) 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 Eq. **12**.

Going back to DP-GD, to circumvent the difficulty in explicitly solving Eq. **12**, we consider

$$\mathrm{d}\hat{\Theta}(t) = -\nabla\mathcal{L}(\hat{\Theta}(t))\mathrm{d}t + \Sigma\,\mathrm{d}B(t), \quad [13]$$

where $\mathcal{L}(\theta)$ is the original (quadratic) training loss and $B(t)$ is the same Wiener process as in Eq. **12**. The solution of the SDE in Eq. **13** is a multidimensional Ornstein–Uhlenbeck (OU) process which admits a closed form (see, e.g., section 4.4.4 in ref. 46). Let us then define

$$\mathcal{C} := \left\{\theta \text{ s.t. } \|\nabla_\theta\ell(\varphi(x_i)^\top\theta - y_i)\|_2 < C_{\text{clip}},\ \forall i \in [n]\right\}, \quad [14]$$

where $[n] = \{1, \dots, n\}$. The set $\mathcal{C}$ contains the parameters such that clipping does not happen, i.e., $\mathcal{L}_{C_{\text{clip}}}(\theta) = \mathcal{L}(\theta)$. If the entire path of $\Theta(t)$ happens in this region (i.e. $\Theta(t) \in \mathcal{C}$ for all $t \in [0, \tau]$), then $\Theta(\tau) = \hat{\Theta}(\tau)$. This is equivalent to

$$\hat{\Theta}(t) \in \mathcal{C}, \text{ for all } t \in [0, \tau], \quad [15]$$

which is an easier event to control, as $\hat{\Theta}(t)$ is an OU process.

4.2. Analysis of Clipping. We show that, by choosing the hyperparameters as in Eq. **8** and setting $\tau = T\eta$, the event in Eq. **15** happens with high probability. To do so, we decompose $\hat{\Theta}(t) = \mathbb{E}_B\left[\hat{\Theta}(t)\right] + \tilde{\Theta}(t) = \hat{\theta}(t) + \tilde{\Theta}(t)$, where $\tilde{\Theta}(t) := \hat{\Theta}(t) - \mathbb{E}_B[\hat{\Theta}(t)]$ and we use that the expectation of an OU process corresponds to the gradient flow $\mathbb{E}_B[\hat{\Theta}(t)] = \hat{\theta}(t)$ of Eq. **9**. Then, the probability of the event in Eq. **15** is lower bounded by the probability that

$$\left|\varphi(x_i)^\top\tilde{\Theta}(t)\right| + \left|\varphi(x_i)^\top\hat{\theta}(t) - y_i\right| \le \frac{C_{\text{clip}}}{2\,\|\varphi(x_i)\|_2}, \quad [16]$$

for all $i \in [n]$ and $t \in [0, \tau]$. As $C_{\text{clip}} = \sqrt{p}\log^2 n$ (Eq. **8**) and $\|\varphi(x_i)\|_2 = \Theta(\sqrt{p})$ with high probability (*SI Appendix*, Eqs. **296** and **297**), Eq. **16** follows from

$$|\varphi(x_i)^\top\hat{\theta}(t) - y_i| = o(\log^2 n), \quad |\varphi(x_i)^\top\tilde{\Theta}(t)| = o(\log^2 n). \quad [17]$$

To obtain the first inequality in Eq. **17**, we show in Lemma 3.1 of *SI Appendix*, section 3, that, jointly for all $i \in [n]$,

$$\sup_{t\in[0,\tau]}\left|\varphi(x_i)^\top\hat{\theta}(t) - y_i\right| = \mathcal{O}(\log n), \quad [18]$$

with high probability. First, we study the stability of GD by proving that $\left\|\theta^* - \theta^*_{-i}\right\|_2 = \tilde{\mathcal{O}}\left(p^{-1/2}\right)$, where θ^*_{-i} is obtained after removing the i-th sample from the training set. Then, we control the entire trajectory of gradient flow for $t \in [0, \tau]$ via i) an explicit computation based on Lie's product formula for the matrix exponential, and ii) a concentration bound over x_i based on Dudley's (chaining tail) inequality.

To obtain the second inequality in Eq. **17**, we show in Lemma 3.2 of *SI Appendix*, section 3 that, jointly for all $i \in [n]$,

$$\sup_{t\in[0,\tau]}\left|\varphi(x_i)^\top\tilde{\Theta}(t)\right| = \mathcal{O}(\log n), \quad [19]$$

with high probability. 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 removes the attractive drift $-\nabla\mathcal{L}(\hat{\Theta}(t))$ from Eq. **13**. Then, Sudakov-Fernique inequality gives that $\mathbb{E}_B\left[\sup_{t\in[0,\tau]}\left|\varphi(x_i)^\top\tilde{\Theta}(t)\right|\right] \le \mathbb{E}_{z_i}\left[\sup_{t\in[0,\tau]}|z_i(t)|\right]$. Since $z_i(t)$ is a Wiener process, the RHS of the previous equation is $\mathcal{O}\left(\Sigma^2\tau\,\|\varphi(x_i)\|_2^2\right)$ via the reflection principle, and this upper bound is of constant order by Eqs. **7** and **8**. An application of the Borell-TIS inequality concludes the argument by giving that, with high probability, $\sup_{t\in[0,\tau]}|\varphi(x_i)^\top\tilde{\Theta}(t)| \le \mathbb{E}_B[\sup_{t\in[0,\tau]}|\varphi(x_i)^\top\tilde{\Theta}(t)|] + \log n$.

4.3. Analysis of Noise and Early Stopping. As $\Theta(\tau) = \hat{\Theta}(\tau)$ with high probability (Eq. **15**), we study the utility of $\Theta(\tau)$ via the closed form of the OU process $\hat{\Theta}(\tau)$. This boils down to controlling the effects of noise and early stopping, which are decoupled by the decomposition $\hat{\Theta}(\tau) = \hat{\theta}(\tau) + \tilde{\Theta}(\tau)$. In fact, $\tilde{\Theta}(\tau)$ is a mean-0 random variable (in the probability space of B) that captures the noise and $\hat{\theta}(\tau)$ is the deterministic component (with respect to B) that captures the early stopping. As $\varphi(x_i)^\top \tilde{\Theta}(\tau)$ is Gaussian with variance increasing linearly in $\|\varphi(x_i)\|_2^2$, τ, and Σ^2, we have that (see Lemma 3.3 of *SI Appendix*, section 3 for details)

$$\mathbb{E}_{x\sim\mathcal{P}_X}\left[\left(\varphi(x)^\top\tilde{\Theta}(\tau)\right)^2\right] = \tilde{\mathcal{O}}\left(\frac{d^2}{\epsilon^2 n^2}\right), \qquad [20]$$

which implies that noise does not damage utility. It remains to show that early stopping is not detrimental. This is proved in Lemma 3.4 of *SI Appendix*, section 3, which informally states that

$$\mathbb{E}_{x\sim\mathcal{P}_X}\left[\left(\varphi(x)^\top\left(\hat{\theta}(\tau) - \theta^*\right)\right)^2\right] = \tilde{\mathcal{O}}\left(\frac{d}{n} + \frac{n}{d^{3/2}}\right). \qquad [21]$$

The idea of the argument is to decompose the LHS of Eq. **21** in two orthogonal subspaces, i.e., $\varphi(x)^\top P_\Lambda(\hat{\theta}(\tau) - \theta^*)$ and $\varphi(x)^\top P_\Lambda^\perp(\hat{\theta}(\tau) - \theta^*)$. Here, $P_\Lambda \in \mathbf{R}^{p\times p}$ is the projector on the space spanned by the top d eigenvectors of $\Phi^\top\Phi$ and $\Phi \in \mathbf{R}^{n\times p}$ is the feature matrix containing $\varphi(x_i)$ in its i-th row. The rationale is that there is a spectral gap between the d-th and the $(d+1)$-th eigenvalue of the kernel $K = \Phi\Phi^\top$, as proved in Lemma 4.5 of *SI Appendix*, section 4. 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 3.4 in *SI Appendix*, section 3. As a consequence of the spectral gap, $\|P_\Lambda(\hat{\theta}(\tau) - \theta^*)\|_2$ is negligible, since in this subspace $\hat{\theta}(\tau)$ is already close to convergence, despite the early stopping. To control the other subspace, we resort to the bounds in Lemmas 4.14 and 4.15 of *SI Appendix*, section 4.

To conclude, denoting by $\hat{\mathcal{R}}$ and $\mathcal{R}^*$ the generalization error of $\hat{\Theta}(\tau)$ and θ^* respectively, Eqs. **20** and **21** guarantee that $|\hat{\mathcal{R}} - \mathcal{R}^*| = \tilde{\mathcal{O}}\left(\frac{d}{n\epsilon} + \sqrt{\frac{d}{n}} + \sqrt{\frac{n}{d^{3/2}}}\right)$. As $\Theta(\tau) = \hat{\Theta}(\tau)$ (due to Eq. **15**), the result of *Theorem* 2 follows.

Data, Materials, and Software Availability. All codes for generating the figures have been deposited in GitHub (https://github.com/simone-bombari/privacy-for-free) (47).

ACKNOWLEDGMENTS. This research was funded in whole, or in part, by the Austrian Science Fund (FWF) Grant number COE 12. For the purpose of open access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission. The authors were also supported by the 2019 Lopez-Loreta prize, and Simone Bombari was supported by a Google PhD fellowship. We thank Diyuan Wu, Edwige Cyffers, Francesco Pedrotti, Inbar Seroussi, Nikita P. Kalinin, Pietro Pelliconi, Roodabeh Safavi, Yizhe Zhu, and Zhichao Wang for helpful discussions.

1. N. Carlini *et al.*, "Extracting training data from large language models" in *USENIX Conference on Security Symposium* (USENIX Association, 2021).
2. N. Haim, G. Vardi, G. Yehudai, O. Shamir, M. Irani, "Reconstructing training data from trained neural networks" in *Advances in Neural Information Processing Systems*, S. Koyejo et al., Eds. (Curran Associates, Inc., 2002).
3. C. Dwork, A. Roth, The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.* **9**, 211–407 (2014).
4. M. Abadi *et al.*, "Deep learning with differential privacy" in *ACM SIGSAC Conference on Computer and Communications Security* (Association for Computing Machinery, 2016).
5. F. Tramer, D. Boneh, "Differentially private learning needs better features (or much more data)" in *International Conference on Learning Representations* (2021).
6. S. De, L. Berrada, J. Hayes, S. L. Smith, B. Balle, Unlocking high-accuracy differentially private image classification through scale. arXiv [Preprint] (2022). http://arxiv.org/abs/2204.13650 (Accessed 16 June 2022).
7. D. Yu, H. Zhang, W. Chen, J. Yin, T. Y. Liu, "Large scale private learning via low-rank reparametrization" in *International Conference on Machine Learning* (PMLR, 2021).
8. A. Golatkar *et al.*, "Mixed differential privacy in computer vision" in *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE Computer Society, 2022).
9. R. Bassily, V. Feldman, K. Talwar, A. Guha Thakurta, "Private stochastic convex optimization with optimal rates" in *Advances in Neural Information Processing Systems*, H. Wallach *et al.*, Eds. (Curran Associates, Inc., 2019).
10. K. Chaudhuri, C. Monteleoni, A. D. Sarwate, Differentially private empirical risk minimization. *J. Mach. Learn. Res.* **12**, 1069–1109 (2011).
11. D. Kifer, A. Smith, A. Thakurta, "Private convex empirical risk minimization and high-dimensional regression" in *Conference on Learning Theory*, S. Mannor, N. Srebro, R. C. Williamson, Eds. (PMLR, 2012).
12. P. Jain, A. G. Thakurta, "(Near) dimension independent risk bounds for differentially" in *International Conference on Machine Learning*, E. P. Xing, T. Jebara, Eds. (PMLR, 2014).
13. R. Bassily, A. Smith, A. Thakurta, "Private empirical risk minimization: Efficient algorithms and tight error bounds" in *2014 IEEE 55th Annual Symposium on Foundations of Computer Science* (IEEE Computer Society, 2014).
14. S. Song, T. Steinke, O. Thakkar, A. Thakurta, "Evading the curse of dimensionality in unconstrained private GLMs" in *International Conference on Artificial Intelligence and Statistics*, A. Banerjee, K. Fukumizu, Eds. (PMLR, 2021).
15. P. Varshney, A. Thakurta, P. Jain, "(Nearly) optimal private linear regression for sub-gaussian data via adaptive clipping" in *Conference on Learning Theory*, P.-L. Loh, M. Raginsky, (PMLR, 2022).
16. D. Yu, H. Zhang, W. Chen, T. Y. Liu, "Do not let privacy overbill utility: Gradient embedding perturbation for private learning" in *International Conference on Learning Representations* (2021).
17. H. Mehta, A. Thakurta, A. Kurakin, A. Cutkosky, Large scale transfer learning for differentially private image classification. arXiv [Preprint] (2022). http://arxiv.org/abs/2205.02973 (Accessed 20 May 2022).
18. Y. Zhou, S. Wu, A. Banerjee, "Bypassing the ambient dimension: Private SGD with gradient subspace identification" in *International Conference on Learning Representations* (2021).
19. H. Asi, J. Duchi, A. Fallah, O. Javidbakht, K. Talwar, "Private adaptive gradient methods for convex optimization" in *International Conference on Machine Learning*, M. Meila, T. Zhang, Eds. (PMLR, 2021).
20. X. Li, F. Tramer, P. Liang, T. Hashimoto, "Large language models can be strong differentially private learners" in *International Conference on Learning Representations* (2022).
21. D. Yu *et al.*, "Differentially private fine-tuning of language models" in *International Conference on Learning Representations* (2022).
22. X. Li *et al.*, "When does differentially private learning not suffer in high dimensions?" in *Advances in Neural Information Processing Systems*, A. H. Oh, A. Agarwal, D. Belgrave, K. Cho, Eds. (Curran Associates Inc., 2022).
23. P. L. Bartlett, A. Montanari, A. Rakhlin, Deep learning: A statistical viewpoint. *Acta Numer.* **30**, 87–201 (2021).
24. P. L. Bartlett, P. M. Long, G. Lugosi, A. Tsigler, Benign overfitting in linear regression. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 30063–30070 (2020).
25. M. Belkin, D. Hsu, S. Ma, S. Mandal, Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 15849–15854 (2019).
26. S. Mei, A. Montanari, The generalization error of random features regression: Precise asymptotics and the double descent curve. *Commun. Pure Appl. Math.* **75**, 667–766 (2022).
27. A. Rahimi, B. Recht, "Random features for large-scale kernel machines" in *Advances in Neural Information Processing Systems*, J. Platt, D. Koller, Y. Singer, S. Roweis, Eds. (Curran Associates, Inc., 2007).
28. S. Mei, T. Misiakiewicz, A. Montanari, Generalization error of random feature and kernel methods: Hypercontractivity and kernel matrix concentration. *Appl. Comput. Harmon. Anal.* **59**, 3–84 (2022).
29. S. Bombari, S. Kiyani, M. Mondelli, "Beyond the universal law of robustness: Sharper laws for random features and neural tangent kernels" in *International Conference on Machine Learning*, A. Krause *et al.*, Eds. (PMLR, 2023).
30. Y. A. Ma, T. V. Marinov, T. Zhang, Dimension independent generalization of DP-SGD for overparameterized smooth convex optimization. arXiv [Preprint] (2022). http://arxiv.org/abs/2206.01836 (Accessed 3 June 2022).
31. R. Arora, R. Bassily, C. A. Guzmán, M. Menart, E. Ullah, "Differentially private generalized linear models revisited" in *Advances in Neural Information Processing Systems*, S. Koyejo *et al.*, Eds. (Curran Associates Inc., 2022).
32. T. T. Cai, Y. Wang, L. Zhang, The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *Ann. Stat.* **49**, 2825–2850 (2021).
33. M. Avella-Medina, C. Bradshaw, P. L. Loh, Differentially private inference via noisy optimization. *Ann. Stat.* **51**, 2067–2092 (2023).
34. X. Liu, P. Jain, W. Kong, S. Oh, A. S. Suggala, Near optimal private and robust linear regression. arXiv [Preprint] (2023). http://arxiv.org/abs/2301.13273 (Accessed 30 January 2023).
35. S. Song, K. Chaudhuri, A. D. Sarwate, "Stochastic gradient descent with differentially private updates" in *2013 IEEE Global Conference on Signal and Information Processing* (IEEE, 2013).
36. M. E. A. Seddik, C. Louart, M. Tamaazousti, R. Couillet, "Random matrix theory proves that deep learning representations of GAN-data behave as gaussian mixtures" in *International Conference on Machine Learning*, H. Daumé III, A. Singh, Eds. (PMLR, 2020).
37. Z. Wang, Y. Zhu, "Overparameterized random feature regression with nearly orthogonal data" in *International Conference on Artificial Intelligence and Statistics*, F. Ruiz, J. Dy, J.-W. van de Meent, Eds. (PMLR, 2023).
38. C. Zhang, S. Bengio, M. Hardt, B. Recht, O. Vinyals, "Understanding deep learning requires rethinking generalization" in *International Conference on Learning Representations* (2017).
39. H. Hu, Y. M. Lu, T. Misiakiewicz, Asymptotics of random feature regression beyond the linear scaling regime. arXiv [Preprint] (2024). http://arxiv.org/abs/2403.08160 (Accessed 13 March 2024).
40. T. Hastie, A. Montanari, S. Rosset, R. J. Tibshirani, Surprises in high-dimensional ridgeless least squares interpolation. *Ann. Stat.* **50**, 949–986 (2022).
41. G. Raskutti, M. J. Wainwright, B. Yu, Early stopping and non-parametric regression: An optimal data-dependent stopping rule. *J. Mach. Learn. Res.* **15**, 335–366 (2014).

42. B. Ghorbani, S. Mei, T. Misiakiewicz, A. Montanari, Linearized two-layers neural networks in high dimension. *Ann. Stat.* **49**, 1029–1054 (2021).
43. D. Wang, C. Chen, J. Xu, "Differentially private empirical risk minimization with non-convex loss functions" in *International Conference on Machine Learning*, K. Chaudhuri, R. Salakhutdinov, Eds. (PMLR, 2019).
44. P. Kloeden, E. Platen, *Numerical Solution of Stochastic Differential Equations, Stochastic Modelling and Applied Probability* (Springer, Berlin Heidelberg, 2011).
45. C. Paquette, E. Paquette, B. Adlam, J. Pennington, "Homogenization of SGD in high-dimensions: Exact dynamics and generalization properties" in *Mathematical Programming* (Springer, 2024), pp. 1–90.
46. C. Gardiner, *Handbook of Stochastic Methods for Physics, Chemistry, and the Natural Sciences, Proceedings in Life Sciences* (Springer-Verlag, 1985).
47. S. Bombari, privacy-for-free. Codes for Privacy for Free in the Overparameterized Regime. GitHub. https://github.com/simone-bombari/privacy-for-free/. Deposited 18 October 2024.