Applied and Computational Harmonic Analysis 77 (2025) 101749

Contents lists available at ScienceDirect

Applied and Computational Harmonic Analysis

journal homepage: www.elsevier.com/locate/acha

Full Length Article

Efficient identification of wide shallow neural networks with biases

Massimo Fornasier [a,*], Timo Klock [b], Marco Mondelli [c], Michael Rauchensteiner [a]

[a] Department of Mathematics, Bolzmannstraße 3, 85748, Garching, Germany
[b] Deeptech Consulting, Oslo, Norway
[c] Institute of Science and Technology (IST) Austria, Am Campus 1, 3400 Klosterneuburg, Austria

ARTICLE INFO

Communicated by Götz Pfander

ABSTRACT

The identification of the parameters of a neural network from finite samples of input-output pairs is often referred to as the *teacher-student model*, and this model has represented a popular framework for understanding training and generalization. Even if the problem is NP-complete in the worst case, a rapidly growing literature – after adding suitable distributional assumptions – has established finite sample identification of two-layer networks with a number of neurons $m = \mathcal{O}(D)$, D being the input dimension. For the range $D < m < D^2$ the problem becomes harder, and truly little is known for networks parametrized by biases as well. This paper fills the gap by providing efficient algorithms and rigorous theoretical guarantees of finite sample identification for such wider shallow networks with biases. Our approach is based on a two-step pipeline: first, we recover the direction of the weights, by exploiting second order information; next, we identify the signs by suitable algebraic evaluations, and we recover the biases by empirical risk minimization via gradient descent. Numerical results demonstrate the effectiveness of our approach.

1. Introduction

Training a neural network is an NP-complete [28,8] and non-convex optimization problem which exhibits spurious and disconnected local minima [5,42,58]. However, highly over-parameterized networks are routinely trained to zero loss and generalize well over unseen data [60]. In an effort to understand these puzzling phenomena, a line of work has focused on the implicit bias of gradient descent methods [3,4,6,35,36,48,56]. Another popular framework for understanding training and generalization is the so-called *teacher-student model* [10,50,43,31,16,17,44,45,61,24,23,21,22,27,32,33,62]. Here, the training data of a so-called *student network* are assumed to be realizable by an unknown *teacher network*, which interpolates them. This model is justified by the wide literature – both classical and more recent – on memorization capacity [14,41,25,34,11,59,53,9], which shows that generic data can be realized by over-parametrized networks. Furthermore, it has also been proved that, in certain settings, small generalization errors necessarily require identification of the parameters [33]. This leads to the fundamental question of understanding whether efficient algorithms exist that ensure the identification of the teacher parameters and consequently the perfect generalization beyond training data. Existing results mostly focus on the identification of the *weights* of shallow (i.e., two-layer) networks with a number of neurons m scaling linearly in the input dimension D (see the related work discussed below). There is also evidence of the average-case hardness of

* Corresponding author.
E-mail addresses: massimo.fornasier@ma.tum.de (M. Fornasier), timo@deeptechconsulting.no (T. Klock), marco.mondelli@ist.ac.at (M. Mondelli), michael.rauchensteiner@ma.tum.de (M. Rauchensteiner).

https://doi.org/10.1016/j.acha.2025.101749
Received 30 April 2023; Received in revised form 30 January 2025; Accepted 9 February 2025

Available online 17 February 2025

the regime $D^{3/2} < m < D^2$, as weight identification can be reduced to tensor decomposition [33]. Instead, the role of biases is often neglected in the literature, although it is well-known that classical universal approximation results, which are the strength of the teacher-student model, do not hold without biases. In fact, as a simple observation, for odd and continuous activations, networks with no biases can only represent functions that are 0 in 0. This distortion at 0 implies failure of any local L^p approximation by continuity. Furthermore, one cannot simply remove the biases by including them in the weights through dimension augmentation and 1-padding, since this would destroy the incoherence of weights needed for their stable identification.

Main contributions In this paper, we give theoretical guarantees on the recovery of both *weights* and *biases* from finite samples in the challenging regime $D < m < D^2$, under mild assumptions on the smoothness of the activation function, incoherence of the weights, and boundedness of the biases. Specifically, the teacher network is given by

$$f : \mathbb{R}^D \to \mathbb{R}, \quad f(x) := \sum_{j=1}^{m} g(\langle w_j, x\rangle + \tau_j), \tag{1}$$

where $w_1, \dots, w_m$ are unit-norm weights and $\tau_1, \dots, \tau_m$ are bounded biases. We propose a two-step parameter recovery pipeline that decouples the learning of the weights from the recovery of the remaining network parameters. In the first step, we use second order information to recover the weights $w_1, \dots, w_m$ up to signs. The method (cf. Section 3) comes with provable guarantees of recovery up to $m \log^2(m) = O(D^2)$ weights, provided that *(i)* the weights are sufficiently incoherent, and *(ii)* second order derivatives of f carry *enough information*. Our approach is based on the observation that $\nabla^2 f(x) = \sum_{j=1}^{m} g^{(2)}(\langle w_j, x\rangle + \tau_j) w_j \otimes w_j \in \mathcal{W} = \operatorname{span}\{w_1 \otimes w_1, \dots, w_m \otimes w_m\}$ and, hence, multiple samples of independent Hessians allow to compute an approximating subspace $\widehat{\mathcal{W}} \approx \mathcal{W}$. The construction of such a subspace is based exclusively on second order information and differs from earlier proposals as by [27], who advocated for the more computationally expensive use of higher order tensor decompositions. The identification of the weights is then performed by projected gradient ascent, the so-called *subspace power method* [20,30,29], seeking for solutions of

$$\max_{u \in \mathbb{S}^{D-1}} \left\| P_{\widehat{\mathcal{W}}}(u \otimes u) \right\|_F^2 \approx \max_{u \in \mathbb{S}^{D-1}} \left\| P_{\mathcal{W}}(u \otimes u) \right\|_F^2 . \tag{2}$$

In the second step (cf. Section 4), we show how to identify the *signs* by suitable algebraic evaluations and the *biases* by empirical risk minimization via gradient descent. Here, we suitably initialize the algorithm and provide convergence guarantees to the ground-truth biases. The convergence proof is based on a linearization argument inspired by the *neural tangent kernel* (NTK) approach. Our theoretical findings are summarized in the following informal statement.

Theorem 1.1 *(Informal). Let f be the shallow network (1) with D inputs and m neurons such that $m \log^2 m = O(D^2)$. Then, for sufficiently large D, there exists a constructive algorithm recovering all weights and shifts of the network with high probability from $O(Dm^2 \log^2 m)$ network queries.*

A few comments on the complexity are in order. For $m = \mathcal{O}(D)$, the proposed pipeline has guaranteed polynomial complexity in D, m. For $D < m < D^2$, while the pipeline is still guaranteed to converge globally, our findings clarify precisely how the hardness of the problem consists in distinguishing local maximizers of (2). This fine geometrical description is novel, and it could pave the way towards a more refined understanding of the hardness of the network identification problem. Furthermore, our numerical experiments (cf. Section 5) consistently show that the network recovery remains surprisingly successful with an experimentally experienced low complexity up to the information-theoretic upper bound $m \approx D^2/2$.[1]

Related work As a neural network is fully determined by a *finite* number of parameters, it is not at all expected to generically require an infinite amount of training samples as in earlier works [49,1,19,54]. This has motivated the rapidly growing literature on the teacher-student model. A popular setup is to minimize the population risk by assuming Gaussian weights: a two-layer ReLU network with a single neuron is considered by [50], a single convolutional filter by [16,10], multiple convolutional filters by [17], and residual networks by [31]. Gradient descent methods have also been widely studied: [44] considers a single ReLU unit; [45] show global convergence for shallow networks with quadratic activations and local convergence for more general activations; gradient descent is combined with an initialization based on tensor decomposition by [62,61,24]. A local convergence analysis for student networks containing at least as many neurons as the teacher is provided by [63]. Let us highlight that these results neglect the role of biases, and the convergence guarantees are either local or, when global, require a number of neurons $m = \mathcal{O}(D)$. Inspired by papers dating back to the 1990s [12,40,13], the works [43,23,22,27,32,33,62] have explored the connection between differentiation of shallow networks and symmetric tensor decompositions. In particular, [27] exploit the third order derivative tensor of the network, whose rank-1 components are made of the weights $w_1, \dots, w_m$. Such a tensor is assumed to fulfill stringent properties (learnability) that would allow for stable algorithmic tensor decomposition and, hence, weight identification. While this preprint contains seminal ideas, it does not give a rigorous justification of the aforementioned stringent properties, nor it provides numerical results showing

[1] This information-theoretic upper bound holds for all methods employing order-2 tensors. In fact, for $m \approx D(D-1)/2$, $\operatorname{span}\{w_1 \otimes w_1, \dots, w_m \otimes w_m\}$ coincides with the space of all symmetric matrices. In particular, in this regime the landscape of the objective function in (2) becomes flatter and flatter, making it impossible to distinguish approximations to $w_j \otimes w_j$ from any other rank-1 matrix competitor.

the success of the proposed tensor decomposition approach. In fact, there is substantial evidence of a computational barrier to solve tensor decomposition in the regime $D^{3/2} < m < D^2$ [33]. In contrast, [23] follow the simpler strategy of the evaluation of the Hessian matrix of the network at different points, yielding a more favorable learnability condition (see **(M3)** below), which can be rigorously justified. Moreover, the results by [23] ensure the provable recovery of weights for $m \leq D$ by a robust matrix optimization promoting minimum rank selection. Once the weights have been identified, the computation of *biases* has been considered by direct estimation [23], or by Fourier methods [27]. To conclude, existing results do not offer rigorous guarantees for the regime $D < m < D^2$.

Technical tools and innovations *Weight identification:* We follow the simple seminal strategy by [23], which exploits the information coming from the Hessians. However, while [23] require the weights to be linearly independent (hence, $m \leq D$), we tackle the challenging case $m > D$. Furthermore, for the identification of the weights we use (2), namely a robust non-linear program over vectors, which is significantly less computationally expensive than the minimum rank selection by [23]. Our analysis further improves upon [20] by allowing to go beyond a linear scaling between m and D and it pushes up to $m = O(D^2)$ by taking advantage of the new insights provided by [29] on the subspace power method.

Shift identification: Differently from [23,27], we set up an empirical risk minimization problem, and we solve it via gradient descent. Our proof of convergence is based on certain kernel matrices, which are reminiscent of those appearing in NTK theory [26]. The NTK perspective has been used to prove global convergence of gradient descent for shallow [18,39,47,57,34,46] and deep neural networks [2,15,64,65,38,37,9]. The technical innovations of our paper with respect to this line of work are as follows. First, we exchange the role between input variable x and weights: we consider the Jacobian of the network with respect to *its input* x, and not to its parameters. This allows us to keep fixed the size of the network and to analyze the NTK spectrum for large input samples. Second, we extend the NTK theory to handle networks with biases. Finally, as the accuracy of the linearization argument depends on the errors accumulated in the weight identification step, we carry out a delicate perturbation analysis.

Notation Given two vectors u and v, let $u \otimes v$ be their Kronecker product and $u \odot v$ their element-wise product. Given a vector u, let $\|u\|_2$ be its ℓ_2 norm and $\mathrm{diag}(v)$ the diagonal matrix with v on diagonal. Given a matrix A, let $\|A\|$ be its operator norm, $\|A\|_F$ its Frobenius norm, and $\|A\|_{F\to F} = \sup_{\|X\|_F=1} \|AX\|_F$. Let $\mathrm{Sym}(\mathbb{R}^{d\times d})$ be the space of symmetric matrices in $\mathbb{R}^{d\times d}$, $C^n(\mathbb{R})$ the space of functions in $\mathbb{R}$ with n continuous derivatives, $\mathrm{Uni}(\mathbb{S}^{D-1})$ the uniform distribution on the D-dimensional sphere $\mathbb{S}^{D-1}$, and Id_p the identity matrix in $\mathbb{R}^{p\times p}$. Given a function g, let $g^{(n)}$ be its n-th derivative. Given a vector v and a permutation π, let v_π be the vector obtained by permuting the entries of v according to π.

2. Network model and main result

We consider the parameter recovery of a shallow neural network f of the form (1). We assume the weights to be drawn uniformly from the sphere, i.e., $w_1, \dots, w_m \sim_{\text{i.i.d.}} \mathrm{Uni}(\mathbb{S}^{D-1})$, and the shifts to be contained in a given interval, i.e., $\tau_1, \dots, \tau_m \in [-\tau_\infty, +\tau_\infty]$. We make the following assumptions on the activation g and on the Hessians of f.

(M1) $g \in C^3(\mathbb{R})$ and

$$\kappa := \max_{n\in[3]} \left\| g^{(n)} \right\|_\infty < \infty. \tag{3}$$

Furthermore, $g^{(2)}$ is strictly monotonic on $(-\tau_\infty, +\tau_\infty)$, $g^{(1)}$ is strictly positive or negative on $(-\tau_\infty, +\tau_\infty)$ and there exists $s \in \{-1, +1\}$ such that for all $\tau \in [-\tau_\infty, +\tau_\infty]$ we have

$$s = \mathrm{sgn}\left(\int_{\mathbb{R}} g^{(1)}(t+\tau) \exp(-t^2/2) dt \right).$$

(M2) $g^{(1)}$ is not a polynomial of degree 3 or less and $\int_{\mathbb{R}} g(t)^2 \exp(-t^2/2) dt < \infty$.

(M3) The Hessians of f have sufficient information for weight recovery, i.e.,

$$\lambda_m \left(\mathbb{E}_{X\sim\mathcal{N}(0,\mathrm{Id})} [\mathrm{vec}(\nabla^2 f(X))^{\otimes 2}] \right) \geq \alpha > 0. \tag{4}$$

The size of the interval $[-\tau_\infty, +\tau_\infty]$ does not depend on m or D, but only on g via **(M1)**. This is satisfied by common activations, e.g., $g(x) = \tanh(x)$ for $\tau_\infty \approx 0.6$ and the sigmoid $g(x) = 1/(1+\exp(-x))$ for $\tau_\infty \approx 1.5$. Condition **(M3)** is common in the related literature [23,21,20], and it guarantees that combining Hessians of f at sufficiently many generic inputs provides enough information to recover all individual weights. A way to show that (4) holds is as follows. First, note that $\nabla^2 f(x) = \sum_{k=1}^m g^{(2)}(w_k^\top x + \tau_k) w_k \otimes w_k \in \mathrm{span}\{w_1 \otimes w_1, \dots, w_m \otimes w_m\}$. Hence, by exploiting the incoherence of $w_1, \dots, w_m \sim \mathrm{Uni}(\mathbb{S}^{D-1})$, one can relate the smallest eigenvalue in (4) to that of the matrix with entries $(\mathbb{E}_{X\sim\mathcal{N}(0,\mathrm{Id})}[g^{(2)}(\langle w_k, X\rangle + \tau_k) g^{(2)}(\langle w_\ell, X\rangle + \tau_\ell)])_{k,\ell}$. This last quantity is then bounded using the tools developed in Section 4.2. This argument can be made rigorous, thus ensuring that **(M3)** holds with $\alpha > 0$ independent of D and m. We also assume the ability to evaluate the network f and to approximate its derivatives.

(G1) We can query the teacher network f and the activation g at any point without noise, and the number of neurons m is known.

Algorithm 1: Network reconstruction.

Input: Teacher neural network f defined in (1) with m neurons, numerical differentiation method $\Delta^n[\cdot]$ with accuracy ϵ, number of Hessian locations N_h and gradient descent samples N_{train}, number of steps for refinement via gradient descent N_{GD}.

1 Compute weights $\widehat{W} = [\hat{w}_1 | \dots | \hat{w}_m]$ by PCA of Hessians followed by iterations of the subspace power method (SPM) from [30] discussed in Section 3;

2 Find signs $\hat{s}$ and initial shifts $\hat{\tau} \in \mathbb{R}^m$ by linearization through higher order differentiation along approximated weight vectors (cf. Algorithm 3 in the supplementary materials, and discussion in Section 4.1);

3 Set $\widehat{W} \leftarrow \widehat{W}\,\mathrm{diag}(\hat{s})$ and construct a student network $\hat{f}$ as in (7) with parameters $\widehat{W}, \hat{\tau}$;

4 Draw samples $x_1, \dots, x_{N_{\text{train}}} \sim \mathcal{N}(0, \mathrm{Id}_D)$ and refine the shifts of $\hat{f}$ by minimizing $J(\hat{\tau})$ (cf. (8)) via gradient descent for N_{GD} steps (cf. Section 4.2). Denote by $\hat{\tau}^{[N_{\text{GD}}]}$ the final iterate.

Output: Weights $\widehat{W}$ and final shifts $\hat{\tau}^{[N_{\text{GD}}]}$ of $\hat{f}$.

(G2) We assume access to a numerical differentiation method, denoted by $\Delta^n[\cdot]$, computing the derivatives for $n = 1, 2, 3$ up to an accuracy $\epsilon > 0$, such that

$$\left\| \nabla^n g(w^\top x) - \Delta^n[g(w^\top x)] \right\|_F \le C_\Delta \left\| w^{\otimes n} \right\|_F \epsilon,$$

where C_Δ is a universal constant only depending on the activation via κ, see (3). Furthermore, for any $b, t_0 \in \mathbb{R}$ the derivatives of $t \mapsto g(bt)$ can be approximated as

$$\left| \frac{d^n}{dt^n} g(b \cdot t)\Big|_{t=t_0} - \Delta^n[g(b\cdot)](t_0) \right| \le C_\Delta b^{n+2} \epsilon. \tag{5}$$

We also assume that the numerical differentiation method is linear, i.e.,

$$\Delta^n[a \cdot g + h] = a \cdot \Delta^n[g] + \Delta^n[h], \tag{6}$$

for any functions g, h and scalar $a \in \mathbb{R}$. Finally, the numerical differentiation algorithm requires a number of queries equal to the dimension of the approximated derivative, i.e., $\mathcal{O}(1)$ for partial derivatives and $\mathcal{O}(D^n)$ for n-th order derivative tensors. Note that all these properties are fulfilled by a standard central finite difference scheme.

Our proposed algorithm for the recovery of the parameters of the planted model (1) is based on a two-step procedure. In the first step, we learn the weight vectors (up to a sign) from the space spanned by Hessian approximations of f (cf. Section 3). Recovering the weights provides access to vectors $\hat{w}_k$, which satisfy $s_k \hat{w}_k \approx w_k$ for some signs $s_1, \dots, s_m \in \{-1, 1\}$. In the second step, we identify the signs $s = (s_1, \dots, s_m)$ and shifts $\tau = (\tau_1, \dots, \tau_m)$ (cf. Section 4). We begin by finding s and an initialization of the shifts $\hat{\tau} \approx \tau$ by a linearization through higher order (numerical) differentiation along the previously computed weight approximations. The shift approximation $\hat{\tau}$ is then refined by empirical risk minimization. More precisely, we consider the parametrization

$$\hat{f}(x, \hat{\tau}) := \sum_{k=1}^{m} g(s_k \langle \hat{w}_k, x \rangle + \hat{\tau}_k), \tag{7}$$

which is fit against the planted model $f(x)$ defined in (1) by minimizing the least squares objective

$$J(\hat{\tau}) = \frac{1}{2N_{\text{train}}} \sum_{i=1}^{N_{\text{train}}} \left(f(x_i) - \hat{f}(x_i, \hat{\tau}) \right)^2 \tag{8}$$

via gradient descent, where $x_1, \dots, x_{N_{\text{train}}} \sim_{\text{i.i.d.}} \mathcal{N}(0, \mathrm{Id}_D)$. Provided that the activation function satisfies **(M1)**-**(M2)**, we show that gradient descent is guaranteed to converge locally to the ground truth shifts up to an error depending only on the accuracy of the initial weight estimates $\hat{w}_k \approx \pm w_k$. The combination of these two steps leads to Algorithm 1 and to our main result, stated below. Its proof is deferred to Appendix D, and it follows as a combination of Theorem 3.3, Proposition 4.2, and Theorem 4.3 (discussed in the rest of the paper).

Theorem 2.1 *(Main result on network reconstruction). Consider the teacher network f defined in (1), where $w_1, \dots, w_m \sim Uni(\mathbb{S}^{D-1})$ and $\tau_1, \dots, \tau_m \in [-\tau_\infty, \tau_\infty]$. Assume g satisfies **(M1)**-**(M2)** and f satisfies the learnability condition **(M3)** for some $\alpha > 0$. Assume we run Algorithm 1 with $N_h > t(m + m^2 \log(m)/D)$ for some $t \ge 1$ and $N_{train} > m\sqrt{D}$. Then, there exists $D_0 \in \mathbb{N}$ and a constant $C > 0$ only depending on g and τ_∞ such that the following holds with probability at least $1 - m^{-1} - 2D^2 \exp(-\min\{\alpha, 1\} t/C) - Cm^2 \exp(-\sqrt{D}/C)$: If $m \ge D \ge D_0$, $Cm \log^2 m \le D^2$, and the numerical differentiation accuracy ϵ satisfies*

$$\epsilon \le \frac{D^{1/2} \min\{1, \alpha^{1/2}\}}{Cm^{9/2} \log(m)^{3/2}}, \tag{9}$$

then Algorithm 1 returns weights and shifts $(\widehat{W} = [\hat{w}_1 | \dots | \hat{w}_m], \hat{\tau}^{[N_{GD}]})$ that fulfill

$$\max_{k \in [m]} \|\hat{w}_{\pi(k)} - w_k\|_2 \le C(m/\alpha)^{1/4} \epsilon^{1/2}, \tag{10}$$

$$\|\hat{\tau}_\pi^{[N_{GD}]} - \tau\|_2 \le C\left(\frac{m^{7/4} D^{1/4} \epsilon^{1/2}}{\alpha^{1/4} N_{train}^{1/2}} + \frac{\xi^{N_{GD}}}{m^{1/2}} + \Delta_{W,1}\right), \tag{11}$$

for some permutation π and some constant $\xi \in [0,1)$ where

$$\Delta_{W,1} := \frac{m^{1/2}\log(m)^{3/4}}{D^{1/4}} \cdot \left(\|\widehat{W} - W\|_F + \frac{\Delta_{W,O}^{1/2}}{D^{1/2}} + \left\|\sum_{k=1}^m (w_k - \hat{w}_k)\right\|_2\right), \tag{12}$$

$$\Delta_{W,O} := \sum_{k\neq k'}^m \left|\langle w_k - \hat{w}_k, w_{k'} - \hat{w}_{k'}\rangle\right|. \tag{13}$$

By choosing an appropriate numerical accuracy ϵ, (9) is satisfied and the error on the weights in (10) can be made arbitrarily small. The error on the shifts in (11) depends on three terms. The first term scales with $\sqrt{\epsilon/N_{\text{train}}}$, hence it is controlled by taking a large number of training samples. The second term vanishes exponentially with the number of gradient steps N_{GD}. Thus, for large enough N_{train} and N_{GD}, the dominant factor is $\Delta_{W,1}$. This last term decreases with the weight approximation error, i.e., if $\widehat{W} = W$, then $\Delta_{W,1} = 0$. In fact, $\Delta_{W,1}$ scales with $\epsilon^{1/2}$, hence it can be reduced by improving the numerical accuracy.

It is natural to compare the residual error term $\Delta_{W,1}$ *after gradient descent* with the error on the shifts *before gradient descent*, i.e., at initialization as given by Proposition 4.2 (cf. (21)). If we assume randomness on the weight errors (with variance matching the upper bound in (10)), i.e., $\hat{w}_{\pi(k)} - w_k \sim_{i.i.d.} \mathcal{N}(0, (m/\alpha)^{(1/2)}\epsilon/D \cdot \mathrm{Id}_D)$ then, up to poly-logarithmic factors, $\Delta_{W,1}$ scales as

$$\frac{\epsilon^{1/2}}{\alpha^{1/4}}\left(\frac{m^{5/4}}{D^{1/4}} + \frac{m^{7/4}}{D}\right). \tag{14}$$

This last quantity is provably smaller than the error (21) at initialization, see the discussion after Proposition 4.2. In the worst case, when all weight errors are aligned, $\Delta_{W,1}$ is dominated by $\left\|\sum_{k=1}^m (w_k - \hat{w}_k)\right\|_2 = \mathcal{O}(m^{5/4}\alpha^{-1/4}\epsilon^{1/2})$, which would not lead to a provable improvement over (21). However, in Section 5, we numerically observe that this type of error accumulation does not occur: the term $\left\|\sum_{k=1}^m (w_k - \hat{w}_k)\right\|_2$ is negligible and $\Delta_{W,1}$ is significantly smaller than (21), see Fig. 2 and the related discussion.

3. Identification of the weights

Definition 3.1 *(RIP).* Let $W \in \mathbb{R}^{D\times m}$, $1 \le p \le m$ be an integer, and $\delta \in (0,1)$. We say that W is (p,δ)-*RIP* if every $D \times p$ submatrix W_p of W satisfies $\|W_p^\top W_p - \mathrm{Id}_p\| \le \delta$.

Definition 3.2 *(Properties of isotropic random weights).* Let $W := [w_1|\dots|w_m]$ and $(G_n)_{k\ell} := \langle w_k, w_\ell\rangle^n$. We define the following incoherence properties:

(A1) There exists $c_1 > 0$, depending only on δ, such that W is $(\lceil c_1 D/\log(m)\rceil, \delta)$-RIP.
(A2) There exists $c_2 > 0$, independent of m, D, so that $\max_{i\neq j}\langle w_i, w_j\rangle^2 \le c_2\log(m)/D$.
(A3) There exists $c_3 > 0$, independent of m, D, so that $\left\|G_n^{-1}\right\| \le c_3$, for all $n \ge 2$.

If the number of weights m is $o(D^2)$, weights drawn from the uniform spherical distribution fulfill **(A1)**-**(A3)** with high probability. This follows from a result due to [29] (cf. Proposition A.1 in Appendix A). We are going to use the properties of Definition 3.2 throughout our analysis.

The weight recovery consists of two steps. First, we leverage the fact that approximated Hessians of the network expose the weights according to

$$\Delta^2 f(x) \approx \nabla^2 f(x) = \sum_{k=1}^m g^{(2)}(\langle w_k, x\rangle + \tau_k) w_k \otimes w_k,$$

such that independent sampling of Hessian locations eventually spans (approximately) the space

$$\widehat{\mathcal{W}} \approx \mathcal{W} := \mathrm{span}\left\{w_1 \otimes w_1, \dots, w_m \otimes w_m\right\}, \tag{15}$$

with $\widehat{\mathcal{W}}, \mathcal{W} \subset \mathrm{Sym}(\mathbb{R}^{D\times D})$. This holds w.h.p. for Hessian locations $x_1, \dots, x_{N_h}$ drawn as standard Gaussians as a consequence of **(M3)**, provided N_h is sufficiently large. The resulting approximation error $\left\|P_{\mathcal{W}} - P_{\widehat{\mathcal{W}}}\right\|_{F\to F}$ can be controlled by the accuracy of the numerical differentiation ϵ, see Lemma A.2 in the supplementary materials. To compute the approximation $\widehat{\mathcal{W}} \approx \mathcal{W}$, one can certainly use finite difference schemes as specified in **(G2)** below, which require actively querying specific points; however, we can also use other passive methods, which do not require querying the network in specific points, but rather in points given by a

distribution p. We refer in particular to the use of weak differentiation and noisy network samples (e.g., with centered and bounded noise) as in formula (4.9) at page 646 of [23]. As in Theorem 4.2 of [23], this formula allows via Algorithm 2 to compute linear combinations of tensors of weights and, w.h.p. and at any accuracy, an approximation of the orthogonal projection onto the subspace $\mathcal{W} = \mathrm{span}(w_1 \otimes w_1, \dots, w_m \otimes w_m)$ with a sample complexity $O(m^2 Q)$, where Q depends on the distribution of the samples (which may in turn depend on the dimension). For instance, one can choose a Gaussian distribution p of the input samples, as it is standard in theoretical machine learning. Hence, the proposed use of noiseless active queries is by no means a restriction, but rather the simplest choice for the purpose of this paper. We refer to [23] for more details on passive sampling.

Next, the weights are uniquely identified (up to a sign) as the $2m$ local maximizers of the program (2), which belong to a certain level set $\{u \in \mathbb{S}^{D-1} \,|\, \left\| P_{\widehat{\mathcal{W}}}(u \otimes u) \right\|_F^2 \geq \beta\}$ of the underlying objective. This follows as a special case from the theory within [30,20,29]. More specifically, [30] study the problem in the unperturbed case, [20] extend the subspace power method to the perturbed objective but their analysis is limited to $m < 2D$, and finally [29] go for 2-tensor decompositions up to $m = o(D^2)$ for the perturbed objective. Then, the local maximizers of (2) are computed via a projected gradient ascent algorithm that iterates

$$u_{j+1} = P_{\mathbb{S}^{D-1}}(u_j + 2\gamma P_{\widehat{\mathcal{W}}}((u_j)^{\otimes 2}) u_j), \tag{16}$$

where γ is the step-size and $P_{\mathbb{S}^{D-1}} / P_{\widehat{\mathcal{W}}}$ denote the projections on $\mathbb{S}^{D-1} / \widehat{\mathcal{W}}$. The iteration (16) starts from a random initialization $u_0 \in \mathbb{S}^{D-1}$, and it was introduced by [30] as a subspace power method (SPM). By iterating (16) until convergence repeatedly from independent starting points, one can collect all m local maximizers of (2) and thereby learn (approximately) all weights up to sign. Assuming the retrieval of every local maximizer is equally likely, the average number of repetitions needed to recover all local maximizers follows from the analysis of the coupon collection problem and grows like $\Theta(m \log m)$ (see also [23]). The theorem below provides a bound on the uniform approximation error for the weights. Its proof, as well as the description of Algorithm 2 summarizing the overall procedure of weight identification, is deferred to Appendix A.

Theorem 3.3 *(Weight recovery). Consider the teacher network f defined in (1), where $w_1, \dots, w_m \sim Uni(\mathbb{S}^{D-1})$ and $\tau_1, \dots, \tau_m \in [-\tau_\infty, \tau_\infty]$. Assume g satisfies **(M1)-(M2)** and f satisfies the learnability condition **(M3)** for some $\alpha > 0$. Then, there exists $D_0 \in \mathbb{N}$ and a constant $C > 0$ depending only on g, τ_∞, such that, for all $D \geq D_0$ and $Cm \log^2 m \leq D^2$, the following holds with probability at least $1 - m^{-1} - D^2 \exp(-\min\{\alpha, 1\} t/C) - C\exp(-\sqrt{m}/C)$: (i) The weights $w_1, \dots, w_m$ fulfill **(A1)-(A3)**, and (ii) if we run Algorithm 2 with numerical differentiation accuracy $\epsilon \leq \frac{\sqrt{\alpha}}{C\sqrt{m}}$ and using $N_h > t(m + m^2 \log(m)/D)$ Hessian locations for some $t \geq 1$, we obtain a set of m approximated weights $\mathcal{U} \subset \mathbb{S}^{D-1}$ such that, for all $k \in [m]$, there exists $\hat{w}_k \in \mathcal{U}$ and a sign $s \in \{-1, +1\}$ for which*

$$\left\| w_k - s\hat{w}_k \right\|_2 \leq C(m/\alpha)^{1/4} \epsilon^{1/2}. \tag{17}$$

4. Identification of the signs and shifts

By leveraging the fact that differentiation exposes the weights of the network as components of the tensor $\nabla^n f(x) = \sum_{k=1}^m g^{(n)}(x^\top w_k + \tau_k) w_k^{\otimes n}$ for $n = 2$, Theorem 3.3 gives that $\hat{w}_k \approx s_k w_k$ for some signs $s_k \in \{-1, +1\}$. In this section, we show how to recover the remaining parameters (shifts and signs) for a given set of ground truth weights $\{w_1, \dots, w_m\} \subset \mathbb{S}^{D-1}$ which are sufficiently incoherent and approximated by $\mathcal{U} = \{\hat{w}_1, \dots, \hat{w}_m\} \subset \mathbb{S}^{D-1}$ up to a sign. This recovery can be broken down into two steps. First, we find the correct signs and good initial shifts (cf. Section 4.1); once the parameters are known, a student network can be initialized from these starting values. Second, the shifts of the student network are refined by empirical risk minimization via gradient descent (cf. Section 4.2).

Remark 4.1. As prefaced in the Theorem 2.1, the recovery of the weights is only possible up to permutations due to the structure of the shallow neural network. Hence, the set of approximated weights lacks any information on what weight approximation belongs to which hidden neuron. One can imagine that we implicitly define the permutation in Theorem 2.1 after recovering the weights by arranging the approximations in $\mathcal{U}$ (cf. Theorem 3.3) in a certain order. Given this order, we then proceed to recover the shifts and signs for a network whose arrangement of hidden neurons matches the inverse of this permutation. To simplify the notation in the following, we assume that this permutation is given by the identity.

4.1. Parameter initialization

Our initialization strategy is centered around the recovery of the quantities $C_2 = (C_{2,1}, \dots, C_{2,m})$ and $C_3 = (C_{3,1}, \dots, C_{3,m})$, where

$$C_{n,k} := s_k^n g^{(n)}(\tau_k), \quad \text{for } k \in [m], \quad n \in \{2, 3\}. \tag{18}$$

If g satisfies **(M1)**, then $g^{(3)}$ does not change sign on the interval $(-\tau_\infty, \tau_\infty)$ due to the monotonicity of $g^{(2)}$. Hence, we can infer the sign s_k from $C_{3,k}$. Furthermore, as $g^{(2)}$ is monotone on $[-\tau_\infty, \tau_\infty]$, it admits an inverse, which allows for the recovery of τ_k from $C_{2,k}$. To learn C_2, C_3, we rely on numerical approximations of the quantities $\langle \nabla^n f(x), \hat{w}_k^{\otimes n} \rangle$, namely, the directional derivatives of the network f along the approximated weights. We consider the following linear system representation of the directional derivatives. Computing the derivative for $x = 0$ reveals

$$\langle \nabla^n f(0), \hat{w}_\ell^{\otimes n} \rangle = \sum_{k=1}^{m} s_k^n g^{(n)}(\tau_k) \langle s_k w_k, \hat{w}_\ell \rangle^n.$$

Denote by $\tilde{G}_n \in \mathbb{R}^{m \times m}$ the matrix with entries $(\tilde{G}_n)_{\ell,k} = \langle \hat{w}_\ell, s_k w_k \rangle^n$. Then, we have

$$\tilde{G}_n \cdot C_n = \begin{bmatrix} \langle \nabla^n f(0), \hat{w}_1^{\otimes n} \rangle \\ \vdots \\ \langle \nabla^n f(0), \hat{w}_m^{\otimes n} \rangle \end{bmatrix} := T_n. \tag{19}$$

In (19), T_n is a vector containing all directional derivatives of f evaluated at 0 along the recovered weights $\hat{w}_1, \dots, \hat{w}_m$. These directional derivatives can be approximated from only $\mathcal{O}(n)$ evaluations of the network by numerical differentiation (cf. **(G2)**), which allows us to compute $\tilde{T}_n \approx T_n$. Provided the weight approximations are sufficiently accurate and incoherent in the sense of Definition 3.2, the matrix $\tilde{G}_n$ is invertible and can be estimated by $(\widehat{G}_n)_{\ell,k} := \langle \hat{w}_\ell, \hat{w}_k \rangle^n$. This follows from Theorem 3.3, which implies that $s_k w_k$ is close to $s_k^2 w_k = w_k$. Therefore, we obtain $C_n \approx \tilde{G}_n^{-1} \tilde{T}_n \approx \widehat{G}_n^{-1} \tilde{T}_n$. This strategy is summarized in Algorithm 3 detailed in Appendix B, and the robustness analysis of Proposition 4.2 makes all the approximations rigorous. This procedure could be carried out for any order of directional derivatives, allowing us to benefit from the higher incoherence of $\langle w_k^{\otimes n}, w_\ell^{\otimes n} \rangle = \langle w_k, w_\ell \rangle^n$. However, for the sake of simplicity and to be more aligned with our network model, we combine only the second and third order directional derivatives.

Proposition 4.2 *(Parameter initialization). Consider the teacher network f defined in (1), where the weights $\{w_k \in \mathbb{S}^{D-1}, k \in [m]\}$ satisfy **(A2)-(A3)** with constants c_2, c_3 and the activation g satisfies **(M1)**. Then, there exist constants $C > 0$ only depending on g, c_2, c_3, τ_∞ and $D_0 \in \mathbb{N}$, such that, for $m \geq D \geq D_0, m \log^2 m \leq D^2$, the following holds. Given $\hat{w}_1, \dots, \hat{w}_m \in \mathbb{S}^{D-1}$ such that*

$$\begin{aligned} \delta_{\max} &:= \max_{k \in [m]} \min_{s \in \{-1,1\}} \left\| w_k - s\hat{w}_k \right\|_2 \\ &\leq \frac{D^{1/2}}{Cm\sqrt{\log m}}, \end{aligned} \tag{20}$$

Algorithm 3 returns a set of shifts $\hat{\tau}$ such that

$$\|\hat{\tau} - \tau\|_2 \leq C\sqrt{m}\epsilon + Cm^{3/2} \left(\frac{\log m}{D} \right)^{3/4} \delta_{\max}, \tag{21}$$

where $\epsilon > 0$ is the accuracy of the numerical differentiation method. Furthermore, once the RHS of (21) is smaller than 1 and $\epsilon \leq (Cm)^{-1}$, the signs returned by Algorithm 3 are identical to the ground truth signs.

The proof is postponed to Appendix B. By Theorem 3.3, we have that $\delta_{\max}$ scales as $(m/\alpha)^{1/4}\epsilon^{1/2}$. Thus, by taking a suitably small ϵ, (20) is satisfied and, after omitting poly-logarithmic factors, the dominant term in (21) scales as

$$\frac{\epsilon^{1/2}}{\alpha^{1/4}} \frac{m^{7/4}}{D^{3/4}}. \tag{22}$$

By comparing (14) and (22) and recalling that m scales at least linearly in D, it is clear that gradient descent improves upon its initialization, under a random model for the weight errors. This improvement is also evident if we evaluate $\Delta_{W,1}$ on the *actual* weights errors coming from the proposed algorithmic pipeline (cf. Fig. 2).

4.2. Local convergence of gradient descent

So far, we have obtained weight approximations $\widehat{W} \approx W$ and shift approximations $\hat{\tau} \approx \tau$ of the shallow teacher network f defined in (1). These parameters $(\widehat{W}, \hat{\tau})$ allow us to define the neural network $\hat{f}$ in (7) and, depending on the accuracy of the previous steps, we would expect already a strong similarity between realizations of $\hat{f}$ and f. In this section, we explore to what degree the approximation $\hat{f}$ can further be improved by tuning the shifts $\hat{\tau}$ in a teacher-student setting. Assume $x_1, \dots, x_{N_{\text{train}}}$ generic inputs and access to N_{train} input-output pairs $(x_i, y_i)_{i \in [N_{\text{train}}]} = (x_i, f(x_i))_{i \in [N_{\text{train}}]}$ of the network f. Based on the initial network configuration of $\hat{f}$, we seek to learn the shifts τ attributed to f by minimizing the least-squares objective (8) via the gradient descent scheme

$$\hat{\tau}^{(n+1)} = \hat{\tau}^{(n)} - \gamma \nabla J(\hat{\tau}^{(n)}). \tag{23}$$

Here, $\gamma > 0$ represent the step-size of the gradient updates. For the case $\widehat{W} = W$, we show that w.h.p. the gradient descent iteration (23) produces a sequence $(\tau^{(n)})_{n \in \mathbb{N}}$ that converges linearly to τ provided that $\|\hat{\tau}^{(0)} - \tau\|_2 = \mathcal{O}(m^{-1/2})$. In the perturbed case where $\widehat{W} \approx W$, we provide an analysis that estimates the error of the shifts w.r.t. *(i)* the Frobenius error $\|\widehat{W} - W\|_F$, *(ii)* the alignment between the individual weight errors $\Delta_{W,O}$ (cf. (13)), and *(iii)* $\| \sum_{k=1}^m w_k - \hat{w}_k \|_2$. More precisely, for sufficiently many training samples N_{train}, the gradient descent iteration will settle within distance $\Delta_{W,1}$ of the optimal solution.

Theorem 4.3 *(Local convergence). Consider the teacher network f defined in (1), with shifts $\tau_1, \dots, \tau_m \in [-\tau_\infty, +\tau_\infty]$ and weights $w_1, \dots, w_m \sim Uni(\mathbb{S}^{D-1})$ that are incoherent according to Definition 3.2. Assume g satisfies (M1)-(M2), and consider the least-squares objective J in (8) constructed with $N_{train} \geq m$ network evaluations $y_1, \dots, y_{N_{train}}$ of f, where $y_i = f(X_i)$ and $X_1, \dots, X_{N_{train}} \sim \mathcal{N}(0, \mathrm{Id}_D)$. Let $\hat{f}$ be parameterized by $\widehat{W}$ and $\hat{\tau}$, as in (7). Then, there exists a constant $C > 0$ depending only on g, τ_∞ and $D_0 > 0$ such that the following holds with probability at least $1 - m\exp(-N_{train}/Cm) - 2m^2\exp(-t/C)$ for $t > 0$: Assume $Cm\log^2 m \leq D^2, m \geq D \geq D_0$ and*

$$\|\tau - \hat{\tau}\|_2 + \Delta_W \leq \frac{1}{C\sqrt{m}}, \tag{24}$$

where $\Delta_W = \Delta_{W,1} + \left(\frac{m^3\delta_{\max}^2 t}{N_{train}}\right)^{1/2}$ and $\Delta_{W,1}$ is given by (12). Then, there exists a $\xi \in [0,1)$, such that the gradient descent iteration (23) with sufficiently small step-size $\gamma > 0$ started from $\hat{\tau}^{(0)} = \hat{\tau}$ satisfies

$$\|\hat{\tau}^{(n)} - \tau\|_2 \leq 2\xi^n\|\hat{\tau}^{(0)} - \tau\|_2 + C(1-\xi^n)\Delta_W. \tag{25}$$

Note that (24) can always be satisfied within our framework as all factors depend on ϵ which can be chosen freely. The proof of Theorem is in Appendix C. The idea is to express J as a quadratic form $J(\hat{\tau}) = (\hat{\tau} - \tau)^\top A(\hat{\tau})(\hat{\tau} - \tau)$, where $A(\hat{\tau})$ denotes the Jacobian obtained by taking derivatives w.r.t. the input features. Then, we linearize around the true solution by replacing $A(\hat{\tau})$ with $A(\tau)$. Analyzing the idealized objective $(\hat{\tau} - \tau)^\top A(\tau)(\hat{\tau} - \tau)$ requires to guarantee the well-posedness of $A(\tau)$, which we prove by using techniques from the NTK literature adapted to our setting (Appendix C.1). The error due to the replacement of $A(\hat{\tau})$ with $A(\tau)$ depends on the error in the weight approximation, and we control it via a delicate argument exploiting Hermite expansions and the incoherence of the weights (Appendix C.2). The decay rate of (25) is largely determined by the factor $\xi \in [0,1)$ which is derived in Lemma C.10.

5. Numerical results

We corroborate our theoretical results by testing the pipeline of Algorithm 1, in order to identify parameters of shallow networks of the type $f(x) = \sum_{k=1}^m \tanh(w_k^\top x + \tau_k)$. As assumed in the theory, the weights are given by $w_1, \dots, w_m \sim_{\text{i.i.d.}} \text{Uni}(\mathbb{S}^{D-1})$, the shifts are sampled according to $\tau_1, \dots, \tau_m \sim_{\text{i.i.d.}} \text{Uni}(-0.5, 0.5)$ and the activation satisfies (M1)-(M2). The number of neurons m depends on the input dimension D according to the rule $m = \lceil \frac{2}{5} D^\beta \rceil$, where the exponent $1/2 \leq \beta \leq 2$ is referred to as the *order of neurons*. The accuracy is evaluated via the following metrics: *(i)* the uniform error (of the approximating network), computed as $E_\infty = m^{-1}\max_i |f(x_i) - \hat{f}(x_i)|$ on a set of 10^6 unseen Gaussian inputs, *(ii)* the worst weight approximation, i.e., $\max_{k\in[m]} \|w_k - \hat{w}_k\|_2$, and *(iii)* the error of the shift approximation, i.e., $m^{-1/2}\|\tau - \hat{\tau}\|_2$. The scaling m^{-1} of E_∞ normalizes for the fact that the range of $f(x)$ grows with m according to our network model (1). All experiments were performed using one NVIDIA Tesla® P100 16GB/GPU in an NVIDIA DGX-1.

Baseline As a baseline for our pipeline, we first try to identify the network parameters in a standard teacher-student setup. The teacher network is fit by empirical risk minimization via SGD applied to a student network of identical architecture. Using 8 minutes of training time with Tensorflow and the hardware as stated above ($N_{\text{train}} = \frac{5}{2} m \cdot D^2$ teacher evaluations, mini-batch size of 64 and learning rate 0.005), we obtain the uniform error depicted in the top row on the left in Fig. 1. These results are averaged over four repetitions. The experiment shows that SGD manages to identify the network parameters and achieve a low uniform error as long as the number of neurons m is small, in particular much smaller than a quadratic scaling such as $m = \lceil \frac{2}{5} D^2 \rceil$. Furthermore, the results worsen for growing dimension D despite higher incoherence of the network weights, possibly due to the fixed training time and learning rate. In an attempt to improve these results, we additionally run SGD for 50 minutes and several different learning rates, fixing the case $D = 50$. The results, shown in the top row on the right in Fig. 1, indicate an improvement of SGD for certain hyperparameter combinations, yet we were not able to find a suitable tuning for $D = 50, m = 1000$. For this experiment, we choose $\tau_1, \dots, \tau_m \sim_{\text{i.i.d.}} \mathcal{N}(0, 0.05)$, thus the ground-truth shifts are closer to the initialization (set at 0) than if they are uniform in $[-0.5, 0.5]$, which should facilitate the task of the SGD algorithm.

Recovery pipeline We now discuss the results of our recovery pipeline in Algorithm 1 to identify shallow networks with tanh activation. For the weight recovery, we use $N_h = \lceil \log(D)m \rceil$ Hessian approximations, which are computed via central finite differences with step-size $\epsilon_{\text{FD}} = 0.01$ and are anchored at evaluations $x_1, \dots, x_{N_h} \sim_{\text{i.i.d.}} \mathcal{N}(0, \mathrm{Id}_D)$. Then, we run $R = 5m\log(m)$ SPM iterations (16) in parallel for 10^3 steps with step-size $\gamma = 2$. The initial shifts computed by the parameter initialization are finalized via (stochastic) gradient descent as described in Section 4. We use $N_{\text{train}} = m \cdot D^2$ samples, learning rate $\gamma = 10^{-3}$ and batch size 64. The training input points are drawn from a standard Gaussian distribution. The refinement of the shifts (by gradient descent) is timed out after 180 seconds, or once we reach a training error below 10^{-8}.

The results of our pipeline in the bottom row of Fig. 1 demonstrate successful recovery of all weights and shifts consistently over 10 repetitions, and for all combinations of m, D. For $\beta = 2$ (or $m = \lceil \frac{2}{5} D^2 \rceil$) the performance of the weight recovery is worse for small D. The causes of this effect may be two-fold: the weights do not yet behave statistically as in the average case scenario for larger D; and the gap between $m = \frac{2}{5} D^2$ and the theoretical limit for weight recovery, $m = D(D-1)/2$, decreases in D. The signs were also recovered successfully. Moreover, we emphasize that the time spent for the weight recovery (which includes the time

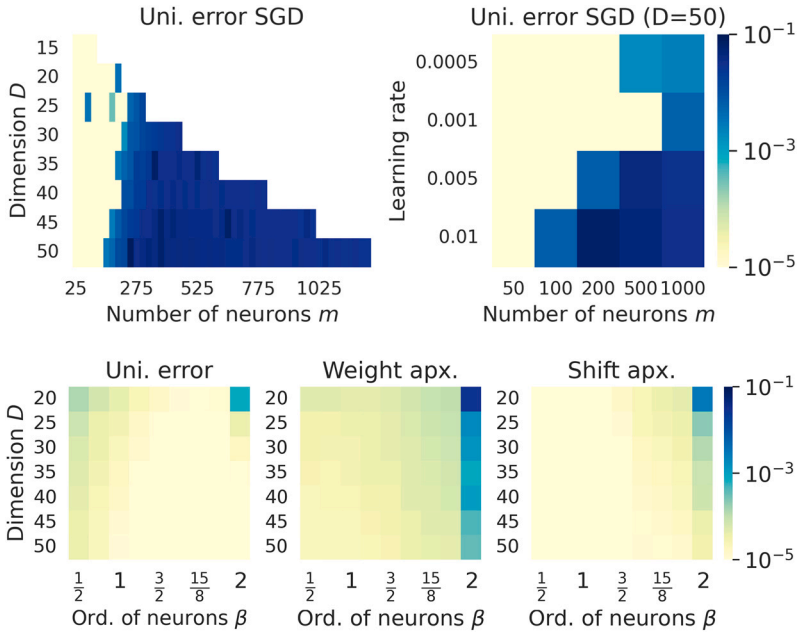

Fig. 1. Performance of parameter identification of shallow networks with tanh activations, m neurons and input size D via SGD for shifts $\tau_k \sim \mathcal{N}(0, 0.05)$ (top row), our pipeline for shifts $\tau_k \sim \mathrm{Uni}(-0.5, 0.5)$ (bottom row).

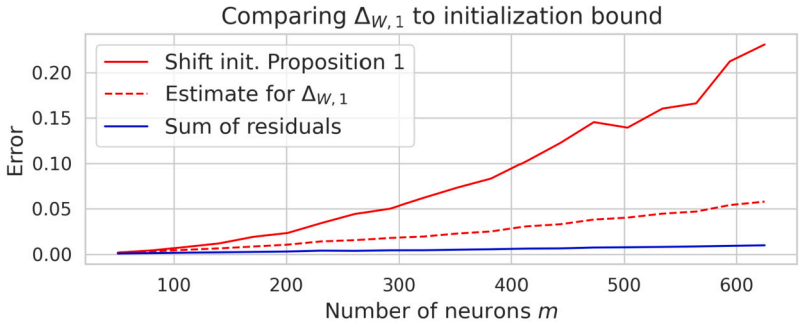

Fig. 2. Comparison between the guaranteed accuracy of the shift initialization (red), the term $\Delta_{W,1}$ (dashed red) and the sum of residual errors $\| \sum_{k=1}^{m} w_k - \hat{w}_k \|_2$ (blue) for weights approximated by our pipeline and $D = 50$.

necessary to approximate all Hessian matrices by numerical differentiation) is in the order of seconds, reaching a maximum of $112s$ for $D = 50, \beta = 2$. The overall runtime of the pipeline is below 5 minutes over all individual runs.

Improvement of the shifts by GD In Fig. 2, we compare the error bound (21) on the initial shifts with $\Delta_{W,1}$ (cf. (12)), where for simplicity the constants C are taken to be 1 in both statements. The results are averaged over 10 realizations. The plot shows that *(i)* the sum of the residuals $\| \sum_{k=1}^{m} w_k - \hat{w}_k \|_2$ in blue has only a negligible contribution to $\Delta_{W,1}$, and hence *(ii)* by settling within distance $\Delta_{W,1}$ of the true shifts, GD will improve over the initialization.

6. Concluding remarks

In this paper, we provide the *first* algorithm with provable guarantees for the finite sample identification of shallow networks *with biases*, where the number of neurons m is roughly $\mathcal{O}(D^2)$. By doing so, we improve upon previous work, which provides guarantees limited to narrow networks (e.g., $m = \mathcal{O}(D)$) or neglects the role of biases. We stress here that our results, beside being rigorous, are also fully numerically reproducible, to show the efficiency of the pipeline. Let us mention that [21,20] have provided partial results on finite sample identification of *deep* networks, yet without rigorous handling of biases. Thus, giving complete guarantees for the case of deep networks, which keep into full consideration also the important role of biases, is an interesting future direction that can build upon the results of the present paper.

Appendix A. Proofs: weight recovery

Algorithm 2 summarizes the first step of the reconstruction pipeline which is the weight recovery. For more details on the exact procedure we refer to Section 3. This section is concerned with the proof of Theorem 3.3, which provides a uniform bound on the approximation error associated with the weight recovery. Additionally, we characterize the incoherence of the resulting approximated weights in terms of the numerical accuracy. A large part of the proofs in this section will operate under the assumptions that vectors, which are drawn uniformly from a high-dimensional sphere, are well separated. To make this more concrete, we rely on a result due to [29] which allows the application of the deterministic incoherence properties **(A1)-(A3)** stated in Definition 3.2 to the ground truth weights $w_1, \dots, w_m \in \mathbb{S}^{D-1}$ which are modeled by a uniform spherical distribution according to our network model (cf. Section 2).

Proposition A.1 *(cf. Proposition 13 in [29]). Let $w_1, \dots, w_m$ be drawn independently from Uni($\mathbb{S}^{D-1}$). If $m = o(D^2)$, then, for any arbitrary constant $\delta \in (0,1)$, there exist constants $C > 0$ and $D_0 \in \mathbb{N}$ depending only on δ such that for all $D \geq D_0$, and with probability at least*

$$1 - m^{-1} - 2\exp(-C\delta^2 D) - C\left(\frac{e \cdot D}{\sqrt{m}}\right)^{-C\sqrt{m}} \tag{26}$$

*conditions **(A1)** - **(A3)** hold with constants $c_2, c_3 < C$.*

Algorithm 2: Weight recovery.

Input: Shallow neural network f, number of neurons m, number of Hessian locations N_h, stepsize $\gamma > 0$, β threshold for rejection of spurious local maximizers

Draw independent samples $x_1, \dots, x_N \sim \mathcal{N}(0, \mathrm{Id})$.

Construct the matrix

$$\widehat{M} := \left[\, \mathrm{vec}(\Delta^2 f(x_1)) \quad \dots \quad \mathrm{vec}(\Delta^2 f(x_N)) \,\right] \in \mathbb{R}^{D^2 \times N_h}.$$

Denote by $P_{\widehat{\mathcal{W}}}$ the orthogonal projection onto the m-th left singular subspace of $\widehat{M}$.

Define $P_{\widehat{\mathcal{W}}}$ as the orth. proj. in matrix space corresponding to $P_{\widehat{\mathcal{W}}}$

Set $\mathcal{U} \leftarrow \emptyset$

while $|\mathcal{U}| < m$ **do**

  Sample $u_0 \sim \mathrm{Unif}(\mathbb{S}^{D-1})$

  Iterate projected gradient ascent

$$u \leftarrow P_{\mathbb{S}^{D-1}}(u + 2\gamma P_{\widehat{\mathcal{W}}}((u)^{\otimes 2})u)$$

  until convergence, and denote the vector of the final iteration by $\hat{u}$.

  if $\left\| P_{\widehat{\mathcal{W}}}(\hat{u}^{\otimes 2}) \right\|_F^2 > \beta$ **then**

    if $\hat{u} \notin \mathcal{U}$ *and* $-\hat{u} \notin \mathcal{U}$ **then**

      $\mathcal{U} \leftarrow \mathcal{U} \cup \{\hat{u}\}$

    end

  end

end

Output: $\mathcal{U}$

Proof sketch of Theorem 3.3 The proof of Theorem 3.3 relies on two individual auxiliary statements. First, a subspace approximation bound covered in Lemma A.2, that controls the error $\left\| P_{\mathcal{W}} - P_{\widehat{\mathcal{W}}} \right\|_{F \to F}$. Recall that $\widehat{\mathcal{W}}$ is constructed to approximate the matrix space

$$\mathcal{W} = \mathrm{span}\{w_1 \otimes w_1, \dots w_m \otimes w_m\} \subset \mathrm{Sym}(\mathbb{R}^{D \times D})$$

from which individual weights can be identified as the rank-1 spanning elements. It is noteworthy that the proof of Lemma A.2 as well as Algorithm 2 makes use of the following convention: We associated every matrix subspace with a classical vector space induced by vectorization. The vectorization of a matrix is denoted by the operator $\mathrm{vec}(\cdot)$ whose output applied to a matrix $X \in \mathbb{R}^{a \times b}$ is the vector in $\mathbb{R}^{a \cdot b}$ containing the columns of X stacked on top of other, i.e.

$$\mathrm{vec}\left(\begin{bmatrix} | & & | \\ x_1 & \dots & x_b \\ | & & | \end{bmatrix}\right) = \begin{pmatrix} x_1 \\ \vdots \\ x_b \end{pmatrix}.$$

This allows us to associate a space like $\mathcal{W} \subset \mathrm{Sym}(\mathbb{R}^{D \times D})$ with the space

$$\mathrm{span}\{\mathrm{vec}(w_1 \otimes w_1), \dots, \mathrm{vec}(w_m \otimes w_m)\} \subset \mathbb{R}^{D^2}.$$

Lemma A.2. *Consider the teacher network f defined in (1). Assume the activation g satisfies **(M1)-(M2)** and f satisfies the learnability condition **(M3)** for some $\alpha > 0$. Furthermore assume that the network weights $w_1, \dots, w_m \in \mathbb{S}^{D-1}$ fulfill **(A2)** of Definition 3.2 with constant*

c_2. Let $P_{\mathcal{W}}$ be the orthogonal approximation onto $\mathcal{W} = \operatorname{span}\{w_1 \otimes w_1, \dots w_m \otimes w_m\}$ and let $P_{\widehat{\mathcal{W}}}$ be constructed as described in Algorithm 2. Then there exists a constant $C > 0$ depending only on g and c_2, such that for numerical diff. accuracy $\epsilon < \frac{\sqrt{\alpha}}{C\sqrt{m}}$ and $N_h > t(m + m^2 \log(m)/D)$ for some $t \geq 1$ we have

$$\left\| P_{\mathcal{W}} - P_{\widehat{\mathcal{W}}} \right\|_{F \to F} \leq C\sqrt{m/\alpha} \cdot \epsilon, \tag{27}$$

with probability at least $1 - D^2 \exp\left(-\frac{t\alpha}{C}\right)$.

Proof. Consider $X_1, \dots, X_{N_h}$ independent copies of a standard Gaussian, i.e. $X_i \sim \mathcal{N}(0, \mathrm{Id}_D)$. Denote by $\mathcal{P}_{\mathcal{W}} \in \mathbb{R}^{D^2 \times D^2}$ the orthogonal projection matrix onto $\operatorname{span}\left\{\operatorname{vec}(w_k \otimes w_k) \,|\, k = 1, \dots, m\right\}$ and by M the matrix with columns given by the exact vectorized Hessians at the inputs $X_1, \dots, X_{N_h}$, i.e.

$$M := \left[\operatorname{vec}(\nabla^2 f(X_1)) \quad \dots \quad \operatorname{vec}(\nabla^2 f(X_{N_h}))\right] \in \mathbb{R}^{D^2 \times N_h}. \tag{28}$$

We associate the matrix subspaces $\mathcal{W}$ and $\widehat{\mathcal{W}}$ with their corresponding D^2-dimensional vector subspaces described by the orthogonal projection matrices $\mathcal{P}_{\mathcal{W}}, \mathcal{P}_{\widehat{\mathcal{W}}}$, respectively. Note that

$$\left\| P_{\mathcal{W}} - P_{\widehat{\mathcal{W}}} \right\|_{F \to F} = \sup_{\|U\|_F = 1} \left\| P_{\mathcal{W}}(U) - P_{\widehat{\mathcal{W}}}(U) \right\|_F = \left\| \mathcal{P}_{\mathcal{W}} - \mathcal{P}_{\widehat{\mathcal{W}}} \right\|$$

with $\|\cdot\|$ describing the ordinary spectral normal in $\mathbb{R}^{D^2}$. Hence, to prove the result, we can rely on the well-known Wedin bound, see for instance [23,21,20,29], giving

$$\left\| P_{\mathcal{W}} - P_{\widehat{\mathcal{W}}} \right\|_{F \to F} = \left\| \mathcal{P}_{\mathcal{W}} - \mathcal{P}_{\widehat{\mathcal{W}}} \right\| \leq \frac{\|M - \widehat{M}\|_F}{\sigma_m(\widehat{M})}, \tag{29}$$

for as long as $\sigma_m(\widehat{M}) > 0$. We continue to provide separate bounds for the numerator and denominator of (29). For the numerator we obtain

$$\begin{aligned}
\|M - \widehat{M}\|_F &\leq \sqrt{N_h} \max_{i \in [N_h]} \left\| \nabla^2 f(X_i) - \Delta^2 f(X_i) \right\|_F \\
&\leq \sqrt{N_h m} \max_{\substack{i \in [N_h] \\ k \in [m]}} \left\| \nabla^2 g(w_k^\top X_i + \tau_k) - \Delta^2 g(w_k^\top X_i + \tau_k) \right\|_F \\
&\leq \max_{k \in [m]} C_\Delta \sqrt{N_h m} \left\| w_k \otimes w_k \right\|_F \epsilon = C_\Delta \sqrt{N_h m} \epsilon,
\end{aligned}$$

where we used the linearity of Δ^2, ∇^2 in the second step and our assumptions on the numerical differentiation method **(G2)** in the last line which gives rise to the constant C_Δ that only depends on g. For the denominator in (29) we use Weyl's inequality [55] which leads to the lower bound

$$\sigma_m(\widehat{M}) \geq \sigma_m(M) - \|M - \widehat{M}\| \geq \sigma_m(M) - \|M - \widehat{M}\|_F \geq \sigma_m(M) - C_\Delta \sqrt{N_h m} \epsilon. \tag{30}$$

Lastly, we need to control $\sigma_m(M)$ by a concentration argument in combination with the learnability Assumption **(M3)** of Section 2. We first express $\sigma_m(M)$ as sum of independent matrices:

$$\sigma_m(M)^2 = \sigma_m(M M^\top) = \sigma_m\left(\sum_{i=1}^{N_h} \operatorname{vec}(\nabla^2 f(X_i)) \otimes \operatorname{vec}(\nabla^2 f(X_i)) \right). \tag{31}$$

Denote $A_i = \operatorname{vec}(\nabla^2 f(X_i)) \otimes \operatorname{vec}(\nabla^2 f(X_i))$. By **(M3)** we know that

$$\sigma_m\left(\sum_{i=1}^{N_h} \mathbb{E} A_i \right) = N_h \alpha > 0.$$

We will make use of the matrix Chernoff (see [51] Corollary 5.2 and the following remark) which states that

$$\mathbb{P}\left(\sigma_m\left(\sum_{i=1}^{N_h} A_i \right) \leq (1-s)\sigma_m\left(\sum_{i=1}^{N_h} \mathbb{E} A_i \right) \right) \leq D^2 \exp\left(-(1-s)^2 \sigma_m\left(\sum_{i=1}^{N_h} \mathbb{E} A_i \right) / 2K \right) \tag{32}$$

for $s \in [0,1]$ and $K = \max_{i \in [N_h]} \|A_i\|_2$. The norm of A_i can be bound uniformly over all $x \in \mathbb{R}^D$ by

$$\left\| \operatorname{vec}(\nabla^2 f(X_i)) \otimes \operatorname{vec}(\nabla^2 f(X_i)) \right\|_2 \leq \sup_{x \in \mathbb{R}^D} \left\| \operatorname{vec}(\nabla^2 f(x)) \right\|_2^2 = \sup_{x \in \mathbb{R}^D} \left\| \nabla^2 f(x) \right\|_F^2$$

$$= \sup_{x\in\mathbb{R}^D} \left\| \sum_{k=1}^{m} g^{(2)}(w_k^\top x + \tau_k) w_k \otimes w_k \right\|_F^2$$

$$= \sup_{x\in\mathbb{R}^D} \sum_{k,\ell=1}^{m} g^{(2)}(w_k^\top x + \tau_k) g^{(2)}(w_\ell^\top x + \tau_\ell) \langle w_k, w_\ell \rangle^2$$

$$\leq \kappa^2 \sum_{k,\ell=1}^{m} \langle w_k, w_\ell \rangle^2 \leq \kappa^2 (m + c_2 m(m-1)\log m/D).$$

The last inequality follows by the incoherence Assumption **(A2)** from the initial statement. Combining this with (32) for $s = 1/2$ together with the bound on the spectrum of the expectation yields

$$\mathbb{P}\left(\sigma_m \left(\sum_{i=1}^{N_h} A_i \right) \geq \frac{1}{2} N_h \alpha \right) \geq 1 - D^2 \exp\left(- \frac{N_h D \alpha}{8\kappa^2 (Dm + c_2 m^2 \log m)} \right). \tag{33}$$

Conditioning on this event, and assuming $\epsilon < \sqrt{\alpha / 8C_\Delta^2 m}$ the initial subspace bound now holds as

$$\left\| P_{\mathcal{W}} - P_{\widehat{\mathcal{W}}} \right\|_{F\to F} \leq \frac{\left\| M - \widehat{M} \right\|_2}{\sigma_m(\widehat{M})} \leq \frac{C_\Delta \sqrt{N_h m} \epsilon}{\sqrt{\frac{1}{2} N_h \alpha} - C_\Delta \sqrt{N_h m} \epsilon} = \frac{C_\Delta \sqrt{m} \cdot \epsilon}{\sqrt{\frac{\alpha}{2}} - C_\Delta \sqrt{m} \cdot \epsilon} \tag{34}$$

$$\leq \frac{\sqrt{2} C_\Delta \sqrt{m} \cdot \epsilon}{\sqrt{\alpha}} \tag{35}$$

with said probability. The final result follows by applying the bound on ϵ onto the denominator. More precisely, we need that $C > 2C_\Delta$ to fulfill (34) and $C > 8\kappa^2 \max\{1, c_2\}$ which implies

$$1 - D^2 \exp\left(- \frac{N_h D \alpha}{8\kappa^2 (Dm + c_2 m^2 \log m)} \right) \leq 1 - D^2 \exp(-t\alpha/C),$$

due to our assumption that $N_h > t(m + m^2 \log(m)/D)$. □

The second part of Algorithm 2 performs projected gradient ascent to find the local maximizers of

$$u \mapsto \left\| P_{\widehat{\mathcal{W}}}(u \otimes u) \right\|_F^2, u \in \mathbb{S}^{D-1}. \tag{36}$$

The landscape for this functional, for $m = o(D^2)$, has been recently analyzed (in particular the properties of its local maximizers) by [29] for the general problem of symmetric tensor decomposition. We now provide one of their main statements adopted to the matrix scenario.

Theorem A.3 *(cf. Theorem 16 in [29]). Let $m, D \in \mathbb{N}$ such that $m \log^2(m) \leq D^2$. Assume $w_1, \dots, w_m$ satisfy **(A1)** - **(A3)** of Definition 3.2 for some $\delta, c_1, c_2, c_3 > 0$. Then there exists δ_0, depending only on c_2, c_3 and D_0, Δ_0, C which depend additionally on c_1, such that if $\delta < \delta_0$, $D > D_0$ and $\left\| P_{\mathcal{W}} - P_{\widehat{\mathcal{W}}} \right\|_{F\to F} \leq \Delta_0$, the program (36) has exactly $2m$ second-order critical points in the superlevel set*

$$\left\{ x \in \mathbb{S}^{D-1} \,\middle|\, \left\| P_{\widehat{\mathcal{W}}}(x \otimes x) \right\|_F^2 \geq Cm \log^2(m)/D^2 + 5 \left\| P_{\mathcal{W}} - P_{\widehat{\mathcal{W}}} \right\|_{F\to F} \right\}. \tag{37}$$

Each of these critical points is a strict local maximizer for $\operatorname{argmax}_{u\in\mathbb{S}^{D-1}} \left\| P_{\widehat{\mathcal{W}}}(u \otimes u) \right\|_F^2$. Furthermore, for each such point x^*, there exists a unique $k \in [m]$ such that

$$\min_{s\in\{-1,1\}} \|x^* - s w_k\|_2 \leq \sqrt{\left\| P_{\mathcal{W}} - P_{\widehat{\mathcal{W}}} \right\|_{F\to F}}. \tag{38}$$

This establishes that the local maximizers of (36) that belong to the superlevel set (37) will be close to one of the weights $w_1, \dots, w_m$ up to sign. The projected gradient ascent iteration in Algorithm 2 converges monotonically to one of the constrained stationary points of (36) as shown by [30]. We are now ready to prove the main result on the weight recovery which relies on the lemma above, Proposition A.1, and the machinery developed by [29] represented by Theorem A.3.

Proof of Theorem 3.3. The weights $w_1, \dots, w_m$ of f are drawn uniformly from the unit sphere. By Proposition A.1, and for any $\delta_0 \in (0,1)$, there exists $D_1 \in \mathbb{N}, C_1 > 0$ depending only on δ_0 such that for all $D \geq D_1$ this set of weights fulfills conditions **(A1)**-**(A3)** of Definition 3.2 with constants $c_2, c_3 < C_1$ and with probability at least

$$1 - m^{-1} - 2\exp(-C_1\delta_0^2 D) - C_1\left(\frac{e\cdot D}{\sqrt{m}}\right)^{-C_1\sqrt{m}}.$$

We condition on this event and denote it by E_1 for the remaining part of the proof. Now, due to the incoherence of the weights and according to our initial assumption which includes $N_h > t(m + m^2\log(m)/D)$, the conditions of Lemma A.2 are met which provides an error bound for the subspace which is constructed in the first part of Algorithm 2, such that

$$\left\|P_{\mathcal{W}} - P_{\widehat{\mathcal{W}}}\right\|_{F\to F} \leq C_2\sqrt{m/\alpha}\cdot\epsilon, \tag{39}$$

with probability at least $1 - D^2\exp\left(-t\alpha/C_2\right)$ for a constant C_2 only depending on g. Denote the event that this subspace bound holds by E_2 and assume it occurs, which only depends on the number of Hessians N_h in relationship to D, m. Note that δ_0 can be freely chosen in $(0,1)$. By Theorem A.3 there exist constants D_2, Δ_0, C_3, such that for $D \geq D_2$ and $\left\|P_{\mathcal{W}} - P_{\widehat{\mathcal{W}}}\right\|_{F\to F} \leq \Delta_0$, the local maximizers of the program $\operatorname{argmax}_{u\in\mathbb{S}^{D-1}}\left\|P_{\widehat{\mathcal{W}}}(u\otimes u)\right\|_F^2$ fulfill

$$\min_{s\in\{-1,1\}}\|x^* - sw_k\|_2 \leq \sqrt{\left\|P_{\mathcal{W}} - P_{\widehat{\mathcal{W}}}\right\|_{F\to F}} \leq \sqrt{C_2\sqrt{m/\alpha}\cdot\epsilon}, \tag{40}$$

as long as they belong to the level set

$$\left\{x\in\mathbb{S}^{D-1}\,\middle|\,\left\|P_{\widehat{\mathcal{W}}}(x\otimes x)\right\|_F^2 \geq C_3 m\log^2(m)/D^2 + 5C_2\sqrt{m/\alpha}\cdot\epsilon\right\}.$$

By iterating projected gradient ascent until convergence, every vector $\hat{u}$ will be one of the these local maximizers. Also note that by construction all vectors returned by Algorithm 2 must have unit norm, hence $\mathcal{U}^* \subset \mathbb{S}^{D-1}$. We need to make sure that level set is not empty, which is guaranteed for $C_3 m\log^2(m)/D^2 \leq \frac{1}{4}$ and $\epsilon \leq \frac{\alpha^{1/2}}{20C_2\sqrt{m}}$ which leads to the threshold

$$C_3 m\log^2(m)/D^2 + 5C_2\sqrt{m/\alpha}\cdot\epsilon \leq \frac{1}{4} + \frac{1}{4} = \frac{1}{2}. \tag{41}$$

Therefore, only considering local maximizers that fulfill $\left\|P_{\widehat{\mathcal{W}}}(x\otimes x)\right\|_F^2 \geq 1/2$ will guarantee that all local maximizers are of the kind which satisfies (40). Before we conclude, there are still some points that need to be addressed. To achieve the bound (40) we had to assume that $\left\|P_{\mathcal{W}} - P_{\widehat{\mathcal{W}}}\right\|_{F\to F} \leq \Delta_0$. This is true due to (39) given the accuracy satisfies $\epsilon \leq \frac{\Delta_0\alpha^{1/2}}{C_3 m^{1/2}}$ which is clearly realizable by our initial assumptions on ϵ, since Δ_0 is independent of m, D. Hence, by further unifying also the constants C_1, C_2, C_3, D_1, D_2, we showed that there exist constants $C > 0, D_0 \in \mathbb{N}$ such that for $D \geq D_0$ and $Cm\log^2 m \leq D^2$ all vectors $u \in \mathcal{U}^*$ returned by Algorithm 2 ran with num. accuracy $\epsilon \leq \frac{\sqrt{\alpha}}{C\sqrt{m}}$ will fulfill the uniform error bound,

$$\min_{s\in\{-1,1\}}\|x^* - s\bar{w}_k\|_2 \leq C(m/\alpha)^{1/4}\epsilon^{1/2}, \tag{42}$$

and this result holds with the combined probability

$$1 - D^2\exp(-t\alpha/C) - m^{-1} - 2\exp(-D/C) - C\left(\frac{e\cdot D}{\sqrt{m}}\right)^{-\sqrt{m}/C}. \quad \square \tag{43}$$

The following short result does show a useful property of the spectrum of higher order Grammians which will prove useful for the upcoming part about parameter initialization.

Lemma A.4 *(Higher order Hadamard products). In the setting of Definition 3.2 we have* $\lambda_{\min}(G_{2+\tilde{n}}) \geq \lambda_{\min}(G_2)$ *and thus in particular* $\left\|G_{2+\tilde{n}}^{-1}\right\|_2 \leq c_3$ *for all* $\tilde{n}\in\mathbb{N}_{\geq 0}$ *as well.*

Proof. For each $n\in\mathbb{N}$ the matrix G_n is a Grammian of the tensors $\{w_1^{\otimes n},\dots,w_m^{\otimes n}\}$ and as such it is a positive semidefinite matrix. Since $\lambda_{\min}(A\odot B) \geq \min_i a_{ii}\lambda_{\min}(B)$ for any pair of positive semidefinite matrices A, B, see Theorem 3 in [7], we thus have

$$\lambda_{\min}(G_{2+\tilde{n}}) = \lambda_{\min}(G_2\odot G_{\tilde{n}}) \geq \lambda_{\min}(G_2)\min_i\langle w_i, w_i\rangle^{\tilde{n}} = \lambda_{\min}(G_2). \quad \square$$

Let us conclude this section with an important auxiliary result. As said before we generally operate in a setting where the ground truth weights are sufficiently incoherent and fulfill **(A1)**-**(A3)** of Definition 3.2. It is clear that these properties translate to accurate approximations of the ground truth weights. The following result makes this explicit alongside with a few other minor technical results which will be used throughout the remaining proofs.

Lemma A.5 *(Incoherence of approximated Weights). Assume the ground truth weights* $\{w_k \in \mathbb{S}^{D-1} | k \in [m]\}$ *fulfill* **(A1)-(A3)** *of Definition 3.2 with constants* c_2, c_3 *and that* $D \leq m$. *Then there exists a constant* $C > 0$ *only depending on* c_2, c_3 *such that for approximations* $\{\hat{w}_k \in \mathbb{S}^{D-1} | k \in [m]\}$ *which satisfy the error bound*

$$\max_{k\in[m]} \min_{s\in\{-1,1\}} \left\| s\hat{w}_k - w_k \right\|_2 = \delta_{\max} \leq \frac{1}{C} \frac{D^{1/2}}{m\sqrt{\log m}} \tag{44}$$

condition **(A2)-(A3)** *of Definition 3.2 holds with constants* $2c_2, 2c_3$. *Furthermore, denote* $\tilde{G}_n \in \mathbb{R}^{m\times m}$ *the matrix with entries* $\tilde{G}_{n,\ell k} = \langle \hat{w}_\ell, s_k w_k \rangle^n$, *where* s_k *are the ground truth signs. Then there exists* D_0 *such that for* $m \geq D \geq D_0$, $n = 2, 3$ *the following holds true:*

(i) For all $k \neq \ell$ *we have* $\langle \hat{w}_k, s_\ell w_\ell \rangle^2 \leq 2c_2 \log(m)/D$
(ii) $\tilde{G}_n$ *is invertible and* $\|\tilde{G}_n^{-1}\| \leq 3c_2$
(iii) Denote by $\tilde{G}_n \in \mathbb{R}^{m\times m}$ *the matrix with entries* $\tilde{G}_{n,\ell k} = \langle \hat{w}_\ell, s_k w_k \rangle^n$, *then*

$$\left\| \tilde{G}_n - \hat{G}_n \right\| \leq Cm \left(\frac{\log m}{D} \right)^{(2n-1)/4} \delta_{\max}. \tag{45}$$

Proof. W.l.o.g. we can assume that C is chosen such that

$$\max_{k\in[m]} \min_{s\in\{-1,1\}} \left\| s\hat{w}_k - w_k \right\|_2 = \delta_{\max} \leq \min \left\{ \frac{1}{8} \left(\frac{c_2 \log m}{D} \right)^{1/2}, \frac{D^{1/2}}{8c_3 m\sqrt{2c_2 \log m}} \right\} \tag{46}$$

holds. We start by showing **(A2)** for the approximated weights. Pick any $k, \ell \in [m], k \neq \ell$. A first observation is that we can disregard the sign that appears in (44) since $\langle \hat{w}_k, \hat{w}_\ell \rangle^2 = \langle -\hat{w}_k, \hat{w}_\ell \rangle^2$. So w.l.o.g. assume that both signs are correct and therefore $\left\| \hat{w}_k - w_k \right\|_2 \leq \delta_{\max}$ and $\left\| \hat{w}_\ell - w_\ell \right\|_2 \leq \delta_{\max}$. Then

$$\begin{aligned}
\langle \hat{w}_k, \hat{w}_\ell \rangle^2 &\leq \left(\left| \langle w_k, w_\ell \rangle \right| + \left| \langle \hat{w}_k - w_k, w_\ell \rangle \right| + \left| \langle w_k, \hat{w}_\ell - w_\ell \rangle \right| + \left| \langle \hat{w}_k - w_k, \hat{w}_\ell - w_\ell \rangle \right| \right)^2 \\
&\leq \left(\left| \langle w_k, w_\ell \rangle \right| + 2\delta_{\max} + \delta_{\max}^2 \right)^2 \leq \left| \langle w_k, w_\ell \rangle \right|^2 + 6\delta_{\max} \left| \langle w_k, w_\ell \rangle \right| + 9\delta_{\max}^2 \\
&\leq \frac{c_2 \log m}{D} + 6 \left(\frac{c_2 \log m}{D} \right)^{1/2} \delta_{\max} + 9\delta_{\max}^2 \\
&\leq \frac{c_2 \log m}{D} + \frac{48+9}{64} \frac{c_2 \log m}{D} \leq \frac{2c_2 \log m}{D},
\end{aligned} \tag{47}$$

which proves that **(A2)** is fulfilled by the approximated weights for a constant $2c_2$. Moving on to **(A3)**, we need to bound the minimal eigenvalue of $\hat{G}_n = (\hat{W}^\top \hat{W})^{\odot n}$ from below. Assuming $\hat{G}_2$ is invertible, we know by Lemma A.4 that

$$\|\hat{G}_n^{-1}\| \leq \|\hat{G}_2^{-1}\| \quad \text{for all} \quad n \geq 2.$$

Thus, it is sufficient to show that **(A3)** holds for the approximated weights for $n = 2$. Denote $G_2 = (W^\top W)^{\odot 2}$. Clearly $G_2, \hat{G}_2$ are symmetric, and since **(A3)** holds for the ground truth weights we know that the minimal eigenvalue of G_2 can be bounded by a constant $\left| \sigma_m(G_2) \right| \geq c_3^{-1}$. Hence, by Weyl's inequality we have

$$\left| \sigma_m(\hat{G}_2) \right| \geq c_3^{-1} - \left\| \hat{G}_2 - G_2 \right\|. \tag{48}$$

Our goal is to find an upper bound the spectral norm on the right hand side. Note that the diagonal of both matrices is identical due to the fact that all columns of $\hat{W}$ and W have unit norm, so we focus on the off diagonal exclusively. Via Gershgorin's circle theorem we attain

$$\begin{aligned}
\left\| \hat{G}_n - G_n \right\| &\leq \max_{k\in[m]} \sum_{\substack{\ell=1 \\ \ell\neq k}}^{m} \left| \langle \hat{w}_k, \hat{w}_\ell \rangle^2 - \langle w_k, w_\ell \rangle^2 \right| \\
&= \max_{k\in[m]} \sum_{\substack{\ell=1 \\ \ell\neq k}}^{m} \left| \langle s_k \hat{w}_k, s_\ell \hat{w}_\ell \rangle^2 - \langle w_k, w_\ell \rangle^2 \right| \\
&\leq \max_{k\in[m]} \sum_{\substack{\ell=1 \\ \ell\neq k}}^{m} \left| \langle s_k \hat{w}_k, s_\ell \hat{w}_\ell \rangle + \langle w_k, w_\ell \rangle \right| \left| \langle s_k \hat{w}_k, s_\ell \hat{w}_\ell \rangle - \langle w_k, w_\ell \rangle \right| \\
&\leq 2 \left(\frac{2c_2 \log m}{D} \right)^{1/2} \max_{k\in[m]} \sum_{\substack{\ell=1 \\ \ell\neq k}}^{m} \left| \langle s_k \hat{w}_k - w_k, s_\ell \hat{w}_\ell \rangle + \langle w_k, s_\ell \hat{w}_\ell - w_\ell \rangle \right|
\end{aligned}$$

$$\leq 4\left(\frac{2c_2\log m}{D}\right)^{1/2} m\cdot\delta_{\max}\leq\frac{1}{2c_3},$$

where we used the fact that **(A2)** holds for the ground truth weights and approximated weights in the penultimate inequality followed by the uniform bound in (44) at the end. We conclude with Weyl's inequality which yields

$$\left|\sigma_m(\widehat{G}_2^{-1})\right|\leq\left|\sigma_1(\widehat{G}_2)\right|^{-1}\leq 2c_3. \tag{49}$$

Hence, the approximated weights fulfill **(A3)** with constant $2c_3$ for $n=2$ which extends to $n\geq 2$ by Lemma A.4. Let us now prove $(i)-(iii)$. The first statement follows directly from our proof of **(A2)** for the approximated weights, since for any $k\neq\ell$ we have

$$\langle\hat{w}_k,s_\ell w_\ell\rangle^2\leq\left(\left|\langle w_\ell,w_k\rangle\right|+\left|\langle\hat{w}_\ell-w_\ell,w_k\rangle\right|\right)^2\leq\left(\left|\langle w_\ell,w_k\rangle\right|+\delta_{\max}\right)^2\leq\frac{2c_2\log m}{D},$$

which follows by the chain of inequalities started in (47). To show (iii) we first split the difference $\tilde{G}_n-\widehat{G}_n=D_n+O_n$ into a diagonal part D_n and an off-diagonal part O_n. We have $\|\tilde{G}_n-\widehat{G}_n\|\leq\|D_n\|+\|O_n\|$, and start by controlling $\|O_n\|$ via Gershgorin's circle theorem:

$$\begin{aligned}\|O_n\|&\leq\max_{\ell\in[m]}\sum_{\substack{k=1\\k\neq\ell}}^{m}\left|\langle\hat{w}_k,\hat{w}_\ell\rangle^n-\langle\hat{w}_k,s_\ell w_\ell\rangle^n\right|\\&\leq\max_{\ell\in[m]}\sum_{\substack{k=1\\k\neq\ell}}^{m}\left|\langle\hat{w}_k,\hat{w}_\ell\rangle-\langle\hat{w}_k,s_\ell w_\ell\rangle\right|\left|\sum_{i=1}^{n}\langle\hat{w}_k,\hat{w}_\ell\rangle^{n-i}\langle\hat{w}_k,s_\ell w_\ell\rangle^{i-1}\right|\\&\leq n\left(\frac{2c_2\log m}{D}\right)^{(n-1)/2}\max_{\ell\in[m]}\sum_{\substack{k=1\\k\neq\ell}}^{m}\left|\langle\hat{w}_k,\hat{w}_\ell-s_\ell w_\ell\rangle\right|.\end{aligned}$$

From here we can slightly improve over Cauchy-Schwarz, and instead use that

$$\sum_{\substack{k=1\\k\neq\ell}}^{m}\left|\langle\hat{w}_k,\hat{w}_\ell-s_\ell w_\ell\rangle\right|\leq\sqrt{m-1}\sqrt{\sum_{\substack{k=1\\k\neq\ell}}^{m}\langle\hat{w}_k,\hat{w}_\ell-s_\ell w_\ell\rangle^2}\leq\sqrt{m}\left\|\widehat{W}\right\|\delta_{\max}.$$

Using $\|\widehat{W}\|=\|\widehat{W}^\top\widehat{W}\|^{1/2}\leq\left(1+m\left(\frac{2c_2\log m}{D}\right)^{1/2}\right)^{1/2}$ we arrive at the following bound for the off-diagonal terms:

$$\begin{aligned}\|O_n\|&\leq n\left(\frac{2c_2\log m}{D}\right)^{(n-1)/2}\sqrt{m}\left(1+m\left(\frac{2c_2\log m}{D}\right)^{1/2}\right)^{1/2}\delta_{\max}\\&\leq Cnm\left(\frac{\log m}{D}\right)^{(n-1)/2}\left(\frac{\log m}{D}\right)^{1/4}\delta_{\max}\leq Cnm\left(\frac{\log m}{D}\right)^{(2n-1)/4}\delta_{\max},\end{aligned}$$

where $C>0$ is an absolute constant only depending on c_2 and $m\geq D$ was used in the second inequality. For the diagonal part we receive

$$\begin{aligned}\|D_n\|&=\max_{\ell\in[m]}\left|1-\left|\langle\hat{w}_\ell,w_\ell\rangle\right|^n\right|\\&=\left|1-\min_{\ell\in[m]}\left|\langle\hat{w}_\ell,w_\ell\rangle\right|^n\right|\\&=\left|1-\left|1-\frac{\max_{\ell\in[m]}\|\hat{w}_\ell-w_\ell\|_2^2}{2}\right|^n\right|\\&\leq\left|1-(1-\delta_{\max}^2/2)^n\right|,\end{aligned}$$

where the equalities are using the fact that $\hat{w}_\ell,w_\ell$ are of unit norm for all $\ell\in[m]$. Hence, we attain overall

$$\left\|\tilde{G}_n-\widehat{G}_n\right\|\leq\left|1-(1-\delta_{\max}^2/2)^n\right|+Cnm\left(\frac{\log m}{D}\right)^{(2n-1)/4}\delta_{\max}.$$

For $n=2,3$ and some constant $C_1>0$ depending only on c_2 this can be further simplified using the bound on $\delta_{\max}$ as

$$\left\|\tilde{G}_n-\widehat{G}_n\right\|\leq\delta_{\max}^2+Cnm\left(\frac{\log m}{D}\right)^{(2n-1)/4}\delta_{\max}$$

Algorithm 3: Parameter Initialization.

Input: Approximated weights $\widehat{W}$, numerical differentiation schema $\Delta^n[\cdot]$ with accuracy $\epsilon > 0$, interval on which $g^{(2)}$ is monotonic $[-\tau_\infty, +\tau_\infty]$

Set $\widehat{G}_2 \leftarrow (\widehat{W}^\top \widehat{W})^{\odot 2}$, $\widehat{G}_3 \leftarrow (\widehat{W}^\top \widehat{W})^{\odot 3}$

for $k = 1, \dots, m$ **do**

  Compute directional derivative approximations $\tilde{T}_{2,k} = \Delta^2[f(\cdot \hat{w}_k)](0), \tilde{T}_{3,k} = \Delta^3[f(\cdot \hat{w}_k)](0)$

end

Set $\tilde{C}_2 \leftarrow \widehat{G}_2^{-1}\tilde{T}_2$, $\tilde{C}_3 \leftarrow \widehat{G}_3^{-1}\tilde{T}_3$

for $k = 1, \dots, m$ **do**

$$\tilde{\tau}_k \leftarrow \begin{cases} (g^{(2)})^{-1}(\tilde{C}_{2,k}), & \text{if } (g^{(2)})^{-1} \text{ is defined for } \tilde{C}_{2,k}, \\ \operatorname{argmin}_{t \in [-\tau_\infty, \tau_\infty]} \left| g^{(2)}(t) - \tilde{C}_{2,k} \right| & \text{else,} \end{cases} \tag{50}$$

$$\tilde{s}_k \leftarrow \operatorname{sign}\left(\tilde{C}_{3,k} \cdot g^{(3)}(0)\right), \tag{51}$$

end

Output: $\tilde{\tau}, \tilde{s}$

$$\leq C_1 m \left(\frac{\log m}{D}\right)^{(2n-1)/4} \delta_{\max},$$

which confirms (iii). To prove (ii) we need to show that $\left\|\tilde{G}_n - \widehat{G}_n\right\| \leq c_4/2$ from which the rest follows as before by Weyl's inequality. We can reuse (iii) in combination with (44) obtaining

$$\left\|\tilde{G}_n - \widehat{G}_n\right\| \leq C_1 m \left(\frac{\log m}{D}\right)^{(2n-1)/4} \delta_{\max} \leq C_2 \left(\frac{\log m}{D}\right)^{1/4},$$

for some constant C_2. Hence (ii) is true for $D \geq D_0$ sufficiently large. □

Appendix B. Proofs: parameter initialization

In this section we prove Proposition 4.2, which assesses the quality of shifts computed by Algorithm 3. These initial shifts will later be used as an initialization for gradient descent (cf. Appendix C).

Proof sketch of Proposition 4.2 As discussed in Section 4.1, goal of Algorithm 3 is to recover the vectors

$$C_2 = g^{(2)}(\tau), \quad \text{and} \quad C_3 = s \odot g^{(3)}(\tau).$$

This recovery is only possible up to approximations $\tilde{C}_2, \tilde{C}_3$ due to perturbations accumulated in the weight recovery and errors caused by the numerical approximation of derivatives. The proof begins with an auxiliary statement, namely Lemma B.1, that develops an upper bound on $\|C_n - \tilde{C}_n\|_2$ $(n = 2, 3)$ assuming that the weight recovery achieved a certain level of accuracy. The proof of Proposition 4.2 will then utilize the properties of the activation function (**(M1)-(M2)**) to show that the shifts τ can be approximated by using the components of $\tilde{C}_2 \approx g^{(2)}(\tau)$, whereas the signs of the original weights are revealed by $\tilde{C}_3 \approx s \odot g^{(3)}(\tau)$.

Lemma B.1. *Denote by $\tilde{C}_n$ the coefficient vectors computed by Algorithm 3 for an input network f with ground truth weights $\{w_k \in \mathbb{S}^{D-1} | k \in [m]\}$ which fulfill **(A2)-(A3)** of Definition 3.2 with constants c_2, c_3 and activation g that fulfills **(M1)**. Then, there exist constants $C > 0$ only depending on g, c_2, c_3 and $D_0 \in \mathbb{N}$, such that for $m \geq D \geq D_0, m \log^2 m \leq D^2$, $n = 2, 3$ and provided approximations $\{\hat{w}_k \in \mathbb{S}^{D-1} | k \in [m]\}$ to the ground truth weights such that*

$$\max_{k \in [m]} \min_{s \in \{-1,1\}} \left\| s\hat{w}_k - w_k \right\|_2 = \delta_{\max} \leq \frac{1}{C} \frac{D^{1/2}}{m\sqrt{\log m}}, \tag{52}$$

we have

$$\left\| \tilde{C}_n - s^n \odot g^{(n)}(\tau) \right\|_2 \leq C\sqrt{m}\epsilon + Cm^{3/2} \left(\frac{\log m}{D}\right)^{(2n-1)/4} \delta_{\max}, \tag{53}$$

where s is the vector storing the true signs that are implied by (52).

Proof of Lemma B.1. Denote as in Algorithm 3 $\tilde{T}_{n,k} = \Delta^n[f(\cdot \hat{w}_k)](0)$ and $T_{n,k} = \langle \nabla^n f(0), \hat{w}_k^{\otimes n} \rangle$. By their definition and the linearity of ∇^n, Δ^n we have

$$\left\| T_n - \tilde{T}_n \right\|_\infty = \sup_{k \in [m]} \left| \langle \nabla^n f(0), \hat{w}_k^{\otimes n} \rangle - \Delta^n[f(\hat{w}_k \cdot)](0) \right| \tag{54}$$

$$\leq \sup_{k \in [m]} \sum_{\ell=1}^{m} \left| \frac{\partial^n}{\partial t^n} g(\langle \hat{w}_k, w_\ell \rangle t + \tau_\ell) \Big|_{t=0} - \Delta^n[g(\langle \hat{w}_k, w_\ell \rangle \cdot + \tau_\ell)](0) \right| \tag{55}$$

$$\leq C_\Delta \epsilon \sup_{k\in[m]} \sum_{\ell=1}^{m} \left|\langle \hat{w}_k, w_\ell\rangle\right|^{n+2} \leq C_\Delta \epsilon \left(1 + m\left(\frac{2c_2 \log m}{D}\right)^{\frac{n+2}{2}}\right), \tag{56}$$

where we used the second point of **(G2)** in the last line followed by the incoherence of the approximated weights **(A2)** established in Lemma A.5. Making use of $D^2 \geq m \log^2 m$, this simplifies to

$$\left\|T_n - \tilde{T}_n\right\|_\infty \leq C_1 \cdot \epsilon,$$

with constant $C_1 = (1+4c_2^2)C_\Delta$ for $n = 2, 3$. Coming back to our initial objective, we can express $s^n \odot g^{(n)}(\tau)$ as the product $s^n \odot g^{(n)}(\tau) = T_n \tilde{G}_n$ where $\tilde{G}_n$ describes the matrix with entries given by $(\tilde{G}_n)_{k\ell} = \langle \hat{w}_k, s_\ell w_\ell\rangle^n$. Note that Algorithm 3 constructs $\tilde{C}_n = \hat{G}_n^{-1}\tilde{T}_n$, where $(\hat{G}_n)_{k\ell} = \langle \hat{w}_k, \hat{w}_\ell\rangle^n$. We can reduce our main statement (53) into separate bounds

$$\begin{aligned}
\left\|\tilde{C}_n - s^n \odot g^{(n)}(\tau)\right\|_2 &= \left\|\hat{G}_n^{-1}\tilde{T}_n - \tilde{G}_n^{-1}T_n\right\|_2 && (57)\\
&\leq \left\|\hat{G}_n^{-1}(T_n - \tilde{T}_n)\right\|_2 + \left\|(\hat{G}_n^{-1} - \tilde{G}_n^{-1})T_n\right\|_2 && (58)\\
&\leq \sqrt{m}\left\|\hat{G}_n^{-1}\right\| \left\|T_n - \tilde{T}_n\right\|_\infty + \left\|(\hat{G}_n^{-1} - \tilde{G}_n^{-1})T_n\right\|_2 && (59)\\
&\leq C_1\sqrt{m}\cdot\epsilon + \left\|(\hat{G}_n^{-1} - \tilde{G}_n^{-1})T_n\right\|_2. && (60)
\end{aligned}$$

To bound $\left\|(\hat{G}_n^{-1} - \tilde{G}_n^{-1})T_n\right\|_2$, we first decompose according to

$$\left\|(\hat{G}_n^{-1} - \tilde{G}_n^{-1})T_n\right\|_2 = \left\|\hat{G}_n^{-1}(\hat{G}_n - \tilde{G}_n)\tilde{G}_n^{-1}T_n\right\|_2 = \left\|\hat{G}_n^{-1}(\hat{G}_n - \tilde{G}_n)(s^n \odot g^{(n)}(\tau))\right\|_2. \tag{61}$$

By invoking Definition **(A3)** again, we continue with

$$\left\|\hat{G}_n^{-1}(\hat{G}_n - \tilde{G}_n)(s^n \odot g^{(n)}(\tau))\right\|_2 \leq 2c_3\left\|\hat{G}_n - \tilde{G}_n\right\| \left\|s^n \odot g^{(n)}(\tau)\right\|_2 \leq 2c_3\kappa\sqrt{m}\left\|\hat{G}_n - \tilde{G}_n\right\|, \tag{62}$$

where we used $\left\|g^{(n)}\right\|_\infty \leq \kappa$. The statement then follows by using inequality (iii) of Lemma A.5 onto $\left\|\hat{G}_n - \tilde{G}_n\right\|$ and unifying the involved constants. □

We are now ready to prove the main result for the parameter initialization.

Proof of Proposition 4.2. First note that due the assumptions made, we can freely apply the results of Lemma A.5 and Lemma B.1. As a consequence the approximated weights considered in the statement of Proposition 4.2 fulfill **(A2)**-**(A3)** of Definition 3.2 with constants derived from the ground truth weights as described in Lemma A.5. We continue with the remark that **(M1)** guarantees the existence of the inverse function $g^{(2)^{-1}}$ on $[-\tau_\infty, \tau_\infty]$ and here we can disregard the signs such that

$$g^{(2)^{-1}}(s^2 \odot g^{(2)}(\tau)) = g^{(2)^{-1}}\left(1 \odot g^{(2)}(\tau)\right) = \tau. \tag{63}$$

While $s^2 \odot g^{(2)}(\tau)$ is not directly available, $\tilde{C}_2$ serves as an approximation $\tilde{C}_2 \approx s^2 \odot g^{(2)}(\tau)$. Fix any $k \in [m]$, and assume that

$$\tilde{C}_{2,k} \in \left[\min_{t\in[-\tau_\infty,+\tau_\infty]} g^{(2)}(t), \max_{t\in[-\tau_\infty,+\tau_\infty]} g^{(2)}(t)\right], \tag{64}$$

then by the mean value theorem

$$\begin{aligned}
\hat{\tau}_k &= g^{(2)^{-1}}(\tilde{C}_{2,k}) = g^{(2)^{-1}}\left(g^{(2)}(\tau_k) + \tilde{C}_{2,k} - g^{(2)}(\tau_k)\right)\\
&= g^{(2)^{-1}}\left(g^{(2)}(\tau_k)\right) + \left(\tilde{C}_{2,k} - g^{(2)}(\tau_k)\right)(g^{(2)^{-1}})'(\xi_k)\\
&= \tau_k + \left(\tilde{C}_{2,k} - g^{(2)}(\tau_k)\right)\frac{1}{g^{(3)}\left(g^{(2)^{-1}}(\xi_k)\right)},
\end{aligned}$$

for some $\xi_k \in \left[\min_{t\in[-\tau_\infty,+\tau_\infty]} g^{(2)}(t), \max_{t\in[-\tau_\infty,+\tau_\infty]} g^{(2)}(t)\right]$. Since $g^{(2)}$ is strictly monotonic on $[-\tau_\infty, \tau_\infty]$ and differentiable, we have

$$\theta := \max_{t\in[-\tau_\infty,\tau_\infty]} \left|g^{(3)}(t)\right| > 0.$$

Hence, we can bound $\left|\frac{1}{g^{(3)}\left(g^{(2)^{-1}}(\xi_k)\right)}\right| \leq \theta^{-1}$ from the outgoing Assumption **(M1)**. Applying Lemma B.1 to bound $\left\|g^{(2)}(\tau) - \tilde{C}_2\right\|_2$ therefore yields

$$\|\hat{\tau} - \tau\|_2 \leq \theta^{-1}\left(C\sqrt{m}\epsilon + Cm^{3/2}\left(\frac{\log m}{D}\right)^{3/4}\delta_{\max}\right). \tag{65}$$

Now, assume there is a $k \in [m]$ such that $\tilde{C}_{2,k}$ does not satisfy (64). By the monotonicity we also know that the maximal and minimal value of $g^{(2)}$ are found exactly on $\pm\tau_\infty$. If $\tilde{C}_{2,k}$ does not lie in the image of $g^{(2)}$ on $[-\tau_\infty, +\tau_\infty]$ it has to exceed one of those. We can assume w.l.o.g. that $\tilde{C}_{2,k} > \max_{t\in[-\tau_\infty,+\tau_\infty]} g^{(2)}(t) = g^{(2)}(\tau_\infty)$. Then,

$$\left|g^{(2)}(\tau_\infty) - g^{(2)}(\tau_k)\right| < \left|\tilde{C}_{2,k} - g^{(2)}(\tau_k)\right|,$$

which shows that $g^{(2)}(\tau_\infty)$ is simply a better estimate of $g^{(2)}(\tau_k)$ than $\tilde{C}_{2,k}$, and $g^{(2)^{-1}}$ is also defined for $g^{(2)}(\tau_\infty)$. Hence, the same error bound as above holds for all $k \in [m]$. The expression in (51) yields the correct sign if $\operatorname{sign}(\tilde{C}_{3,k}) = \operatorname{sign}(s_k^{(3)}) \cdot \operatorname{sign}(g^{(3)}(\tau_k)) = \operatorname{sign}(s_k) \cdot \operatorname{sign}(g^{(3)}(\tau_k))$. This is the case if

$$\left|s_k^{(3)} \cdot g^{(3)}(\tau_k)\right| > \left|s_k^{(3)} \cdot g^{(3)}(\tau_k) - \tilde{C}_{3,k}\right|. \tag{66}$$

By our outgoing assumption $\left|s_k^{(3)} \cdot g^{(3)}(\tau_k)\right| \geq \theta$ and together with Lemma B.1 applied to the RHS of the inequality above, we get that the signs are correct as long as

$$\theta > \left(C\sqrt{m}\epsilon + Cm^{3/2}\left(\frac{\log m}{D}\right)^{5/4}\delta_{\max}\right). \tag{67}$$

Assume now that the RHS of (21) is smaller than 1 and $\epsilon \leq (Cm)^{-1}$, this implies in particular

$$Cm^{3/2}\left(\frac{\log m}{D}\right)^{3/4}\delta_{\max} < 1.$$

We can estimate the right hand side of (67) from above by

$$C\sqrt{m}\epsilon + Cm^{3/2}\left(\frac{\log m}{D}\right)^{5/4}\delta_{\max} \leq \frac{1}{m^{1/2}} + \left(\frac{\log m}{D}\right)^{2/4},$$

which clearly is smaller than any constant for D large enough, and therefore the signs will be correct for D_0 chosen accordingly since (66) is fulfilled. □

Appendix C. Proof of Theorem 4.3

Let us shortly recall the setting of Theorem 4.3. We consider the identification of the parameters W, τ attributed to a shallow neural network $f(\cdot, W, \tau)$ which falls into the class of networks described in Section 2. By means of Algorithms 2-3, we can find weight approximations $\widehat{W} \approx W$ and shift approximations $\hat{\tau} \approx \tau$ of f. The parameters $(\widehat{W}, \hat{\tau})$ give rise to a neural network $\hat{f}(\cdot, \widehat{W}, \hat{\tau})$ which is architecturally identical to f, and, depending on the accuracy of the previous algorithmic steps, we would already expect some agreement in terms of $\hat{f} \approx f$. Given network evaluations $y_1 = f(x_1), \ldots, y_{N_{\text{train}}} = f(x_{N_{\text{train}}})$ of f, we consider further improvement of the approximated shifts $\hat{\tau}$ by empirical risk minimization of the objective

$$J(\hat{\tau}) = \frac{1}{2N_{\text{train}}}\sum_{i=1}^{N_{\text{train}}}\left(\hat{f}(x_i, \hat{\tau}) - y_i\right)^2, \tag{68}$$

via gradient descent given by

$$\hat{\tau}^{(n+1)} = \hat{\tau}^{(n)} - \gamma\nabla J(\hat{\tau}^{(n)}). \tag{69}$$

In Theorem 4.3, we prove a local convergence result with the guarantee that, for sufficiently large N_{train}, $\|\hat{\tau}^{(n)} - \tau\|_2$ is roughly

$$\|\hat{\tau}^{(n)} - \tau\|_2 \lesssim \frac{m^{1/2}\log(m)^{3/4}}{D^{1/4}}\left(\|\widehat{W} - W\|_F + \frac{\Delta_{W,O}^{1/2}}{D^{1/2}} + \left\|\sum_{k=1}^m w_k - \hat{w}_k\right\|_2\right),$$

where

$$\Delta_{W,O} = \sum_{k\neq k'}^m \left|\langle \hat{w}_k - w_k, \hat{w}_{k'} - w_{k'}\rangle\right|.$$

Proof sketch For the proof, we rely on an idealized loss given by a quadratic functional in $\hat{\tau}$:

$$J_*(\hat{\tau}) = (\hat{\tau} - \tau)^\top A(\hat{\tau} - \tau), \tag{70}$$

with

$$A := \frac{1}{2N_{\text{train}}}\sum_{i=1}^{N_{\text{train}}}\nabla\hat{f}(x_i, \tau)\nabla\hat{f}(x_i, \tau)^\top. \tag{71}$$

The proof can then be broken down into two steps. First, in Lemma C.4, it is shown that J_* is strictly convex by estimating a lower bound on the minimal eigenvalue $\lambda_m(A)$ of A. The proof relies on techniques from the NTK literature to first control the spectrum of $\mathbb{E}_{X_1,\dots,X_{N_{\text{train}}}\sim\mathcal{N}(0,\mathrm{Id}_D)}[A]$ by leveraging **(M2)** and the incoherence of $\dot{w}_1,\dots,\dot{w}_m$. In particular, Lemma C.4 implies that minimizing J_* via the gradient descent iteration given by

$$\hat{\tau}_*^{(n+1)} = \hat{\tau}_*^{(n)} - \gamma \nabla J_*(\hat{\tau}_*^{(n)}) = \hat{\tau}_*^{(n)} - \gamma A(\hat{\tau}_*^{(n)} - \tau) \tag{72}$$

with step-sizes $\gamma \le 1/\|A\|$ does necessarily converge to the global minimum attained at $\hat{\tau}_* = \tau$. As a second step, we control the perturbation between the iterations $\hat{\tau}^{(n)}, \hat{\tau}_*^{(n)}$, when starting them from an identical vector $\hat{\tau}^{(0)} = \hat{\tau}_*^{(0)}$. In particular, in Lemma C.10 it is shown that the difference $\|\hat{\tau}^{(n)} - \hat{\tau}_*^{(n)}\|_2$ adheres to

$$\|\hat{\tau}^{(n)} - \hat{\tau}_*^{(n)}\|_2 \le \xi^n \|\hat{\tau}^{(0)} - \tau\|_2 + (1-\xi^n)\Delta_W,$$

provided $\hat{\tau}^{(0)}$ is sufficiently close to the optimal solution τ, and where Δ_W is an error term which satisfies $\Delta_W \to 0$ as $\|\widehat{W} - W\|_F \to 0$ and $\xi \in [0,1)$. By the triangle inequality, we then bound the distance of the original gradient descent iteration (69) to τ via

$$\begin{aligned}\|\hat{\tau}^{(n)} - \tau\|_2 &\le \|\hat{\tau}^{(n)} - \hat{\tau}_*^{(n)}\|_2 + \|\hat{\tau}_*^{(n)} - \tau\|_2 \\ &\le \xi^n \|\hat{\tau}^{(0)} - \tau\|_2 + (1-\xi^n)\Delta_W + (1-\gamma\lambda_m(A))^n \|\hat{\tau}_*^{(0)} - \tau\|_2 \to \Delta_W,\end{aligned}$$

for $n \to \infty$. Hence, we establish that the iteration $\hat{\tau}^{(n)}$ settles in an area around the optimal shifts τ that is determined by the initial and irreparable error present in the weight approximation $\widehat{W}$ of W.

Organisation of this section Subsection C.1 is dedicated to analyze the matrix A in (71) in expectation (over x_i's) and proves the well-posedness. Subsection C.2 analyzes the perturbation between gradient descent on the idealized objective J_* and the true objective J. Subsection C.3 concludes the proof by combining the well-posedness and the perturbation analysis.

C.1. Well-posedness of the idealized objective in expectation

We begin this section with a short primer on Hermitian expansions, a technical tool which is commonly used in the NTK literature. Afterwards, we prove the well-posedness of A in (71) in expectation.

C.1.1. A primer on Hermitian expansions

The Hermitian polynomials form an orthonormal basis of the L_2 space, weighted by the Gaussian kernel w_G, which we denote as $L_2(\mathbb{R}, w_G)$. The r-th Hermitian polynomial is defined as

$$h_r(y) := \frac{1}{\sqrt{r!}}(-1)^r \exp\left(\frac{y^2}{2}\right)\frac{d^r}{dy^r}\exp\left(\frac{-y^2}{2}\right).$$

Any function $h \in L_2(\mathbb{R}, w_G)$ can be expanded as $h \equiv \sum_r \mu_r(h) h_r$ with Hermitian coefficients $\mu_r(h)$ as

$$\mu_r(h) := \int h(y) h_r(y) w_G(y) dy.$$

As per Assumption **(M1)** the first three derivatives of g are bounded, hence $\max_{k\in[3]} \|g_\tau^{(k)}\|_\infty < \infty$ for any $\tau \in \mathbb{R}$. It is easy to check that this implies that these functions lie within $L_2(\mathbb{R}, w_G)$.

Lemma C.1. *Assume h is bounded, then $h \in L_2(\mathbb{R}, \omega_G)$ and*

$$\sum_{r\ge 0} \mu_r(h)^2 \le \sqrt{2\pi}\|h\|_\infty^2$$

Proof.

$$\int_{\mathbb{R}} h(t)^2 \exp(-t^2/2) dt \le \sqrt{2\pi}\|h\|_\infty^2 < \infty.$$

The second statement follows from the fact that $L_2(\mathbb{R}, \omega_G)$ is a Hilbert space and the hermite polynomials form an orthonormal system within that space. □

We further assume in **(M2)** that $g^{(1)}$ is not a polynomial of degree three or less, implying that also $g_\tau^{(1)}$ is not a polynomial of degree three or less. Since h_0, h_1, h_2, h_3 form a basis for the space of affine functions, this implies $g_\tau^{(1)} \notin \mathrm{Span}(h_0, h_1, h_2, h_3)$. In particular, $\mu_r(g_\tau^{(1)}) \neq 0$ for some $r \ge 4$ and any $\tau \in \mathbb{R}$. In the following, we denote

$$\omega := \min_{\tau\in[-\tau_\infty,\tau_\infty]} \sum_{r\ge 4} \mu_r(g^{(1)}(\cdot+\tau))^2 > 0,$$

which depends only on the activation function $g^{(1)}$ and the shift bound τ_∞. Lastly, a useful property of Hermitian expansions and the Hermitian basis is the following identity.

Lemma C.2 *(cf. Lemma D.2 in [38]). For two unit norm vectors $x, y \in \mathbb{R}^D$ and every $k, \ell \geq 0$ we have*

$$\mathbb{E}_{X\sim\mathcal{N}(0,\mathrm{Id}_D)}\left[h_k(v^\top X)h_\ell(u^\top X)\right] = \delta_{k\ell}\langle u, v\rangle^k,$$

where $\delta_{k\ell} = 1$ if $k = \ell$ and 0 otherwise.

C.1.2. Well-posedness in expectation

The central object of study in this section is the matrix

$$E := \mathbb{E}_{X_1,\dots,X_{N_{\text{train}}}\sim\mathcal{N}(0,\mathrm{Id}_D)}[A]. \tag{73}$$

We prove its well-posedness in Lemma C.4. The proof relies on the observation that E is actually equal to a sum of positive semidefinite Grammian matrices as shown in Lemma C.3.

Lemma C.3. *Assume that **(M1)** holds, and let E be defined as in (73). Then, we have*

$$E = \frac{1}{2}\sum_{r=0}^{\infty} T_r T_r^\top, \quad \text{where } T_r := \begin{bmatrix} \mu_r(g_{\tau_1}^{(1)})\,\mathrm{vec}(\hat{w}_1^{\otimes r}) \\ \vdots \\ \mu_r(g_{\tau_m}^{(1)})\,\mathrm{vec}(\hat{w}_m^{\otimes r}) \end{bmatrix} \in \mathbb{R}^{m\times D^r}.$$

In particular, we have $E \succcurlyeq \frac{1}{2}\sum_{r\in\mathcal{R}} T_r T_r^\top$ for any subset $\mathcal{R} \subseteq \mathbb{N}_{\geq 1}$, where $A \succcurlyeq B$ means $A - B$ is positive semidefinite.

Proof. The matrix A can be written as

$$A_{k\ell} = \frac{1}{2N_{\text{train}}}\sum_{i=1}^{N_{\text{train}}} g^{(1)}(\hat{w}_k^\top X_i + \tau_k)g^{(1)}(\hat{w}_\ell^\top X_i + \tau_\ell)$$

and the corresponding expectation reads

$$E_{k\ell} = \frac{1}{2}\mathbb{E}_{X\sim\mathcal{N}(0,\mathrm{Id}_D)}\left[g_{\tau_k}^{(1)}(\hat{w}_k^\top X)g_{\tau_\ell}^{(1)}(\hat{w}_\ell^\top X)\right].$$

Now, note that $g_\tau^{(1)} = g^{(1)}(\cdot + \tau) \in L_2(\mathbb{R}, w_H)$ for any $\tau \in \mathbb{R}$ by **(M1)** and Lemma C.1. Hence, $g_\tau^{(1)}$ has a Hermitian expansion and we can write

$$E_{k\ell} = \frac{1}{2}\mathbb{E}_{X\sim\mathcal{N}(0,\mathrm{Id}_D)}\left[\left(\sum_{r=0}^{\infty}\mu_r(g_{\tau_k}^{(1)})h_r(\hat{w}_k^\top X)\right)\left(\sum_{r=0}^{\infty}\mu_r(g_{\tau_\ell}^{(1)})h_r(\hat{w}_\ell^\top X)\right)\right].$$

Using now Lemma C.2 to express expectations of scalar products of Hermite polynomials, we obtain

$$E_{k\ell} = \frac{1}{2}\sum_{r=0}^{\infty}\mu_r(g_{\tau_k}^{(1)})\mu_r(g_{\tau_\ell}^{(1)})\langle\hat{w}_k, \hat{w}_\ell\rangle^r,$$

which can equivalently be written as $\frac{1}{2}\sum_{r=0}^{\infty} T_r T_r^\top$. The second part of the statement follows from the fact that each individual matrix $T_r T_r^\top$ is a positive semidefinite Grammian matrix. □

Lemma C.4. *Let E be defined as in (73) and assume that the approximated weights satisfy $\|\hat{w}_k\|_2 = 1$ and **(A2)** for some universal constant c_2. Furthermore, assume the activation function adheres to **(M2)**. Then, we have*

$$\lambda_m(E) \geq \omega - C(m-1)\left(\frac{\log m}{D}\right)^2, \tag{74}$$

where ω and C are constants depending only on g and τ_∞. Specifically, we have

$$\omega = \frac{1}{2}\min_{\tau\in[-\bar{\tau}_\infty,\bar{\tau}_\infty]}\sum_{r\geq 4}\left(\mu_r(g^{(1)}(\cdot+\tau))\right)^2,$$

$$C = \frac{1}{2}c_2^2\max_{\tau,\widetilde{\tau}\in[-\tau_\infty,\tau_\infty]}\sum_{r\geq 4}\left|\mu_r(g_\tau^{(1)})\mu_r(g_{\widetilde{\tau}}^{(1)})\right|.$$

Proof of Lemma C.4. To simplify the notation, we introduce the shorthand $\mu_{r,k} := \mu_r(g_{\tau_k}^{(1)})$. By Lemma C.3 we have $E \succcurlyeq \frac{1}{2}\sum_{r\geq 4} T_r T_r^\top$, so we concentrate on the expression on the right hand side. As $\|\hat{w}_k\|_2 = 1$ for all $k \in [m]$, we first note that we can rewrite $\frac{1}{2}\sum_{r\geq 4} T_r T_r^\top$ as $\frac{1}{2}\sum_{r\geq 4} T_r T_r^\top = D_4 + O_4$, where the matrix D_4 is given by

$$D_4 := \frac{1}{2}\operatorname{Diag}\Big(\sum_{r\geq 4}\mu_{r,1}^2, \dots, \sum_{r\geq 4}\mu_{r,m}^2\Big)$$

and the remainder O_4 equals $\frac{1}{2}\sum_{r\in 4} T_r T_r^\top$ with its diagonal set to 0. To show (74), we compute a lower eigenvalue bound for D_4 and an upper eigenvalue bound for O_4 independently, and then complete the argument with Weyl's eigenvalue perturbation bound [55]. The smallest eigenvalue of D_4 can be read from the diagonal and is given by

$$\lambda_{\min}(D_4) = \frac{1}{2}\min_{k\in[m]}\sum_{r\geq 4}\mu_{r,k}^2 \geq \omega > 0.$$

For the spectral norm of O_4 we use L_1/L_∞-Cauchy-Schwarz inequalities and $\|\hat{w}_k\|_2 = 1$ for all $k \in [m]$. Specifically, for any unit norm vector u we have

$$\begin{aligned} u^\top O_4 u &= \frac{1}{2}\sum_{k=1}^m \sum_{\ell\neq k} u_k u_\ell \sum_{r\geq 4}\mu_{r,k}\mu_{r,\ell}\langle \hat{w}_k, \hat{w}_\ell\rangle^r \\ &\leq \frac{1}{2}\sum_{k=1}^m \sum_{\ell\neq k} |u_k|\,|u_\ell| \sum_{r\geq 4}|\mu_{r,k}\mu_{r,\ell}|\,|\langle \hat{w}_k, \hat{w}_\ell\rangle|^r. \end{aligned}$$

By dragging out the maximum of the sums over Hermitian coefficients, we further bound

$$u^\top O_4 u \leq \Big(\frac{1}{2}\max_{\tau,\widetilde{\tau}\in[-\tau_\infty,\tau_\infty]}\sum_{r\geq 4}\Big|\mu_r(g_\tau^{(1)})\mu_r(g_{\widetilde{\tau}}^{(1)})\Big|\Big)\sum_{k=1}^m\sum_{\ell\neq k}|u_k|\,|u_\ell|\,|\langle \hat{w}_k, \hat{w}_\ell\rangle|^4.$$

The trailing factor is, for all unit norm u, bounded by the spectral norm of the matrix

$$(\widehat{O}_4)_{ij} := \begin{cases} 0, & \text{if } i = j, \\ \big|\langle \hat{w}_i, \hat{w}_j\rangle\big|^4, & \text{else}. \end{cases} \tag{75}$$

Therefore we have $u^\top O_4 u \leq C_{g,\tau_\infty}\|\widehat{O}_4\|$ for all unit norm u, and with the constant C_{g,τ_∞} given as

$$C_{g,\tau_\infty} = \frac{1}{2}\max_{\tau,\widetilde{\tau}\in[-\tau_\infty,\tau_\infty]}\sum_{r\geq 4}\Big|\mu_r(g_\tau^{(1)})\mu_r(g_{\widetilde{\tau}}^{(1)})\Big|,$$

and only dependent on g and the shift bound τ_∞. By Gershgorin's circle theorem we further have

$$\|\widehat{O}_4\| \leq \max_{k\in[m]}\sum_{\ell\neq k}^m |(\widehat{O}_4)_{k\ell}| \leq (m-1)\left(\frac{c_2\log m}{D}\right)^2,$$

where we used the fact that $\hat{w}_1, \dots, \hat{w}_m$ satisfy **(A2)**. □

C.2. Controlling the perturbation from the idealized GD iteration

This section is concerned with the divergence between the two gradient descent iterations in (69) and (72). We start with a number of auxiliary results that control certain series involving the Hermite coefficients of the activation and its derivatives. These technical statements are needed to control the perturbation in the GD iteration that is caused by the weight approximation. The bounds enable Lemma C.9 which provides an upper bound for the difference between the gradients $\nabla J(\hat{\tau}), \nabla J_*(\hat{\tau})$, defined in (68), (70), respectively, w.r.t. the accuracy of the estimated weights and shift initializations.

C.2.1. Controlling perturbation from weights

The first part of this section is concerned with estimating a series that contains elements

$$S_{r,\ell} := \sum_{k=1}^m \mu_r(g_{\tau_k})\mu_r(g_{\hat{\tau}_\ell}^{(1)})\left(\langle \hat{w}_k, \hat{w}_\ell\rangle^r - \langle w_k, \hat{w}_\ell\rangle^r\right), \tag{76}$$

where $\mu_r(g_{\tau_k}), \mu_r(g_{\hat{\tau}_\ell}^{(1)})$ correspond to the k-th and ℓ-th Hermite coefficient of the function $g_{\tau_k}(\cdot) = g(\cdot + \tau_k)$, $g'_{\hat{\tau}_\ell}(\cdot) = g^{(1)}(\cdot + \hat{\tau}_\ell)$, respectively. These coefficients are assumed to be uniformly bounded for all $r \geq 0$ which is a consequence of **(M2)** and Lemma C.1. The following results pave the way for perturbation bound w.r.t. estimated weights and we use the following shorthands to keep the expressions more compact:

$$\Delta_{W,F} = \|\widehat{W} - W\|_F, \tag{77}$$

$$\Delta_{W,O} = \sum_{k\neq k'}^{m} \left|\langle \hat{w}_k - w_k, \hat{w}_{k'} - w_{k'}\rangle\right|. \tag{78}$$

Lemma C.5. *Consider weights and approximated weights $(w_k)_{k\in[m]}, (\hat{w}_k)_{k\in[m]}$ of unit norm as before that both fulfill (A2) and (i) in Lemma A.5, as well shifts $(\tau_k)_{k\in[m]}, (\hat{\tau}_k)_{k\in[m]}$ within $[-\tau_\infty, \tau_\infty]$ for some $\tau_\infty < \infty$. Let $S_{r,\ell}$ be defined as in (76) and assume that g fulfills the Assumption (M1)-(M2). Then, there exists a constant $C > 0$ such that, for $m \geq D$,*

$$\sum_{\ell=1}^{m} S_{r,\ell}^2 \leq Cr^2 \max_{k,\ell\in[m]} \mu_r(g_{\hat{\tau}_\ell}^{(1)})^2 \mu_r(g_{\tau_k})^2 \left(1 + m\left(\frac{\log m}{D}\right)^{r/2}\right) \cdot \left[\Delta_{W,F}^2 + \left(\frac{\log m}{D}\right)^{(r-1)/2} \Delta_{W,O}\right].$$

Furthermore, for any fixed $R \geq 2$ we have

$$\sum_{r=2}^{R} 2^r \sum_{\ell=1}^{m} S_{r,\ell}^2 \leq \frac{Cm\log m}{D}\left(\Delta_{W,F}^2 + \left(\frac{\log m}{D}\right)^{1/2} \Delta_{W,O}\right),$$

where the constant $C > 0$ additionally depends on R.

Proof of Lemma C.5. Throughout this proof we use the convention that, for any vector, we have $v^{\otimes 0} = 1$, $1 \otimes 1 = 1$ and $v \otimes 1 = 1 \otimes v = v$, which will be relevant for the case $r = 1$. We start with a chain of equalities that uses elementary properties of the Frobenius inner product:

$$\begin{aligned}
\sum_{\ell=1}^{m} S_{r,\ell}^2 &= \sum_{\ell=1}^{m}\left[\sum_{k=1}^{m} \mu_r(g_{\tau_k})\mu_r(g_{\hat{\tau}_\ell}^{(1)})\left(\langle \hat{w}_k, \hat{w}_\ell\rangle^r - \langle w_k, \hat{w}_\ell\rangle^r\right)\right]^2 \\
&= \sum_{\ell=1}^{m}\left[\sum_{k=1}^{m} \mu_r(g_{\tau_k})\mu_r(g_{\hat{\tau}_\ell}^{(1)})\langle \hat{w}_k - w_k, \hat{w}_\ell\rangle\left(\sum_{i=1}^{r}\langle \hat{w}_k, \hat{w}_\ell\rangle^{r-i}\langle w_k, \hat{w}_\ell\rangle^{i-1}\right)\right]^2 \\
&= \sum_{\ell=1}^{m}\left[\sum_{k=1}^{m} \mu_r(g_{\tau_k})\mu_r(g_{\hat{\tau}_\ell}^{(1)})\langle \hat{w}_k - w_k, \hat{w}_\ell\rangle\left\langle \sum_{i=1}^{r} \hat{w}_k^{\otimes(r-i)} \otimes w_k^{\otimes(i-1)}, \hat{w}_\ell^{\otimes r-1}\right\rangle\right]^2 \\
&= \sum_{\ell=1}^{m}\left[\sum_{k=1}^{m} \mu_r(g_{\tau_k})\mu_r(g_{\hat{\tau}_\ell}^{(1)})\left\langle (\hat{w}_k - w_k) \otimes \sum_{i=1}^{r}\left(\hat{w}_k^{\otimes(r-i)} \otimes w_k^{\otimes(i-1)}\right), \hat{w}_\ell^{\otimes r}\right\rangle\right]^2 \\
&= \sum_{\ell=1}^{m}\left[\mu_r(g_{\hat{\tau}_\ell}^{(1)})\left\langle \sum_{k=1}^{m} \mu_r(g_{\tau_k})(\hat{w}_k - w_k) \otimes \sum_{i=1}^{r}\left(\hat{w}_k^{\otimes(r-i)} \otimes w_k^{\otimes(i-1)}\right), \hat{w}_\ell^{\otimes r}\right\rangle\right]^2 \\
&= \sum_{\ell=1}^{m} \mu_r(g_{\hat{\tau}_\ell}^{(1)})^2\left\langle \sum_{k=1}^{m} \mu_r(g_{\tau_k})(\hat{w}_k - w_k) \otimes \sum_{i=1}^{r}\left(\hat{w}_k^{\otimes(r-i)} \otimes w_k^{\otimes(i-1)}\right), \hat{w}_\ell^{\otimes r}\right\rangle^2.
\end{aligned}$$

At this stage, we separate the coefficients depending on ℓ such that

$$\begin{aligned}
&\sum_{\ell=1}^{m}\left[\sum_{k=1}^{m} \mu_r(g_{\tau_k})\mu_r(g_{\hat{\tau}_\ell}^{(1)})\left(\langle \hat{w}_k, \hat{w}_\ell\rangle^r - \langle w_k, \hat{w}_\ell\rangle^r\right)\right]^2 \\
&\leq \max_{\ell\in[m]} \mu_r(g_{\hat{\tau}_\ell}^{(1)})^2 \sum_{\ell=1}^{m}\left\langle \sum_{k=1}^{m} \mu_r(g_{\tau_k})(\hat{w}_k - w_k) \otimes \sum_{i=1}^{r}\left(\hat{w}_k^{\otimes(r-i)} \otimes w_k^{\otimes(i-1)}\right), \hat{w}_\ell^{\otimes r}\right\rangle^2.
\end{aligned}$$

Now, note that the set of tensors $(\hat{w}_\ell^{\otimes r})_{\ell\in[m]}$ forms a frame whose upper frame constant is bounded by the upper spectrum of the Grammian $(\widehat{G}_r)_{ij} = \langle \hat{w}_i, \hat{w}_j\rangle^r$, see also Lemma E.2. Due to Lemma E.3 which relies on Gershgorin's circle theorem we know there exists an absolute constant $C > 0$ such that for D sufficiently large the operator norm of $\widehat{G}_r$ obeys

$$\|\widehat{G}_r\| \leq C\left(1 + m\left(\frac{\log m}{D}\right)^{r/2}\right).$$

As a consequence, we get

$$\begin{aligned}
&\max_{\ell\in[m]} \mu_r(g^{(1)}_{\hat{\tau}_\ell})^2 \sum_{\ell=1}^{m} \left\langle \sum_{k=1}^{m} \mu_r(g_{\tau_k})(\hat{w}_k - w_k) \otimes \sum_{i=1}^{r} \left(\hat{w}_k^{\otimes(r-i)} \otimes w_k^{\otimes(i-1)} \right), \hat{w}_\ell^{\otimes r} \right\rangle^2 \\
&\leq \max_{\ell\in[m]} \mu_r(g^{(1)}_{\hat{\tau}_\ell})^2 \|\widehat{G}_r\| \left\| \sum_{k=1}^{m} \mu_r(g_{\tau_k})(\hat{w}_k - w_k) \otimes \sum_{i=1}^{r} \left(\hat{w}_k^{\otimes(r-i)} \otimes w_k^{\otimes(i-1)} \right) \right\|_F^2 \\
&\leq C \max_{\ell\in[m]} \mu_r(g^{(1)}_{\hat{\tau}_\ell})^2 \left(1 + m \left(\frac{\log m}{D} \right)^{r/2} \right) \left\| \sum_{k=1}^{m} \mu_r(g_{\tau_k})(\hat{w}_k - w_k) \otimes \sum_{i=1}^{r} \left(\hat{w}_k^{\otimes(r-i)} \otimes w_k^{\otimes(i-1)} \right) \right\|_F^2 .
\end{aligned} \tag{79}$$

Denote now $\Delta_{k,r} := \mu_r(g_{\tau_k})(\hat{w}_k - w_k)$ and $T_{k,r} := \sum_{i=1}^{r} \left(\hat{w}_k^{\otimes(r-i)} \otimes w_k^{\otimes(i-1)} \right)$, then

$$\begin{aligned}
&\left\| \sum_{k=1}^{m} \mu_r(g_{\tau_k})(\hat{w}_k - w_k) \otimes \sum_{i=1}^{r} \left(\hat{w}_k^{\otimes(r-i)} \otimes w_k^{\otimes(i-1)} \right) \right\|_F^2 \\
&= \sum_{k,k'=1}^{m} \langle \Delta_{k,r} \otimes T_{k,r}, \Delta_{k',r} \otimes T_{k',r} \rangle = \sum_{k,k'=1}^{m} \langle \Delta_{k,r}, \Delta_{k',r} \rangle \langle T_{k,r}, T_{k',r} \rangle \\
&= \sum_{k=1}^{m} \|\Delta_{k,r}\|_2^2 \|T_{k,r}\|_F^2 + \sum_{k\neq k'}^{m} \langle \Delta_{k,r}, \Delta_{k',r} \rangle \langle T_{k,r}, T_{k',r} \rangle .
\end{aligned} \tag{80}$$

Using $\|w_k\|_2 = \|\hat{w}_k\|_2 = 1$ we get

$$\|T_{k,r}\|_F \leq \sum_{i=1}^{r} \|\hat{w}_k^{\otimes(r-i)} \otimes w_k^{\otimes(i-1)}\|_F \leq \sum_{i=1}^{r} \|\hat{w}_k\|_2^{r-i} \|w_k\|_2^{i-1} = r,$$

such that the left part of (80) can be estimated by

$$\begin{aligned}
\sum_{k=1}^{m} \|\Delta_{k,r}\|_2^2 \|T_{k,r}\|_F^2 &\leq r^2 \max_{k\in[m]} \mu_r(g_{\tau_k})^2 \sum_{k=1}^{m} \|\hat{w}_k - w_k\|_2^2 \\
&= r^2 \max_{k\in[m]} \mu_r(g_{\tau_k})^2 \|\widehat{W} - W\|_F^2 .
\end{aligned} \tag{81}$$

To bound the right part of (80) first note that, for $k \neq k'$,

$$\begin{aligned}
\langle T_{k,r}, T_{k',r} \rangle &= \sum_{i,i'=1}^{r} \left\langle \hat{w}_k^{\otimes(r-i)} \otimes w_k^{\otimes(i-1)}, \hat{w}_{k'}^{\otimes(r-i')} \otimes w_{k'}^{\otimes(i'-1)} \right\rangle \\
&\leq C \sum_{i,i'=1}^{r} \left(\frac{\log m}{D} \right)^{(r-1)/2} = C r^2 \left(\frac{\log m}{D} \right)^{(r-1)/2} ,
\end{aligned}$$

for some absolute constant C, which follows from the pairwise incoherence **(A2)** as well as point (i) of Lemma A.5. Therefore, the right part of (80) is bounded by

$$\begin{aligned}
\sum_{k\neq k'}^{m} \langle \Delta_{k,r}, \Delta_{k',r} \rangle \langle T_{k,r}, T_{k',r} \rangle &\leq C r^2 \left(\frac{\log m}{D} \right)^{(r-1)/2} \sum_{k\neq k'}^{m} \left| \langle \Delta_{k,r}, \Delta_{k',r} \rangle \right| \\
&\leq C r^2 \left(\frac{\log m}{D} \right)^{(r-1)/2} \max_{k\in[m]} \mu_r(g_{\tau_k})^2 \sum_{k\neq k'}^{m} \left| \langle \hat{w}_k - w_k, \hat{w}_{k'} - w'_k \rangle \right| .
\end{aligned} \tag{82}$$

Plugging (81) and (82) into (80) yields

$$\left\| \sum_{k=1}^{m} \mu_r(g_{\tau_k})(\hat{w}_k - w_k) \otimes \sum_{i=1}^{r} \left(\hat{w}_k^{\otimes(r-i)} \otimes w_k^{\otimes(i-1)} \right) \right\|_F^2 \tag{83}$$

$$\leq C r^2 \max_{k\in[m]} \mu_r(g_{\tau_k})^2 \left[\|\widehat{W} - W\|_F^2 + \left(\frac{\log m}{D} \right)^{(r-1)/2} \sum_{k\neq k'}^{m} \left| \langle \hat{w}_k - w_k, \hat{w}_{k'} - w'_k \rangle \right| \right] . \tag{84}$$

Combining this with (79) yields the desired first statement

$$\begin{aligned}
\sum_{\ell=1}^{m} S_{r,\ell}^2 \leq C r^2 &\max_{k,\ell\in[m]} \mu_r(g^{(1)}_{\hat{\tau}_\ell})^2 \mu_r(g_{\tau_k})^2 \left(1 + m \left(\frac{\log m}{D} \right)^{r/2} \right) \\
&\cdot \left[\Delta_{W,F}^2 + \left(\frac{\log m}{D} \right)^{(r-1)/2} \Delta_{W,O} \right] .
\end{aligned}$$

For the second statement, note that $\max_{k\in[m]} \mu_r(g_{\tau_k})^2$ is bounded due to **(M2)** and $\max_{\ell\in[m]} \mu_r(g^{(1)}_{\hat{\tau}_\ell})^2$ is bounded according to Lemma C.1. Hence it follows that

$$\sum_{r=2}^{R} 2^r \sum_{\ell=1}^{m} S_{r,\ell}^2 \leq \sum_{r=2}^{R} 2^r C r^2 \left(1 + m\left(\frac{\log m}{D}\right)^{r/2}\right)\left[\Delta_{W,F}^2 + \left(\frac{\log m}{D}\right)^{(r-1)/2}\Delta_{W,O}\right]$$
$$\leq \left(1 + m\left(\frac{\log m}{D}\right)\right)\left(\Delta_{W,F}^2 + \left(\frac{\log m}{D}\right)^{1/2}\Delta_{W,O}\right)\sum_{r=1}^{R} 2^r C r^2.$$

The second statement follows from the upper bound above by adjusting the constant C due to $\sum_{r=1}^{R} 2^r C r^2 < \infty$ for fixed R and using $(m\log m)/D > 1$. □

Lemma C.6. *Consider weights and approximated weights $(w_k)_{k\in[m]}, (\hat{w}_k)_{k\in[m]}$ of unit norm as before that both fulfill **(A2)** and (i) in Lemma A.5, as well shifts $(\tau_k)_{k\in[m]}, (\hat{\tau}_k)_{k\in[m]}$ within $[-\tau_\infty, \tau_\infty]$ for some $\tau_\infty < \infty$. Let $S_{r,\ell}$ be defined as in (76) and assume that g fulfills the Assumption **(M1)-(M2)**. Then, there exists a constant $C > 0$ such that for $m \geq D$*

$$\sum_{\ell=1}^{m} S_{1,\ell}^2 \leq Cm\left(\frac{\log m}{D}\right)^{1/2}\left\|\sum_{k=1}^{m} w_k - \hat{w}_k\right\|_2^2.$$

Proof. According to the proof of Lemma C.5, in particular (79), we can bound

$$\sum_{\ell=1}^{m} S_{1,\ell}^2 \leq Cm\left(\frac{\log m}{D}\right)^{1/2}\left\|\sum_{k=1}^{m} \mu_1(g_{\tau_k})(w_k - \hat{w}_k)\right\|_2^2,$$

for some constant $C > 0$. Since $\mu_1(g_{\tau_k})$ is bounded for all $k \in [m]$, what remains is to show that the Hermite coefficients do not change signs. Note that the first Hermite polynomial is given by $h_1(u) = u$. According to the definition of the Hermite coefficients we have

$$\mu_1(g_{\tau_k}) = \int_{\mathbb{R}} u g(u+\tau_k) e^{-u^2/2} du = \left[-g(u+\tau_k)e^{-u^2/2}\right]_{-\infty}^{\infty} + \int_{\mathbb{R}} g^{(1)}(u+\tau_k)e^{-u^2/2}du$$
$$= \int_{\mathbb{R}} g^{(1)}(u+\tau_k)e^{-u^2/2}du.$$

Now note that $g^{(1)}(u+\tau_k)$ will always have the same sign since $g^{(2)}$ is monotonic due to **(M1)**. Therefore $\mu_1(g_{\tau_1}), \ldots, \mu_1(g_{\tau_m})$ must all be either positive or negative, from which the proof follows directly. □

Lemma C.7. *Assume that g fulfills the Assumption **(M1)-(M2)** and that the shifts $(\tau_k)_{k\in[m]}, (\hat{\tau}_k)_{k\in[m]}$ are within $[-\tau_\infty, \tau_\infty]$. Then, for $R \geq 4$, we have*

$$\sum_{r\geq R} r \left|\max_{k,\ell\in[m]} \mu_r(g_{\tau_k})\mu_r(g^{(1)}_{\hat{\tau}_\ell})\right| < \infty. \tag{85}$$

Proof. By applying Lemma E.1 (whose condition is met due to **(M1)-(M2)**), we immediately get that, for all $r \geq R$,

$$\mu_r(g_{\tau_k})\mu_r(g^{(1)}_{\hat{\tau}_\ell}) = \left(3!\binom{r}{r-3}\right)^{-1/2}\mu_{r-3}(g^{(3)}_{\tau_k})\cdot\left(2!\binom{r}{r-2}\right)^{-1/2}\mu_{r-2}(g^{(3)}_{\hat{\tau}_\ell})$$
$$= \left((r-2)(r-1)^2r^2\right)^{-1/2}\mu_{r-3}(g^{(3)}_{\tau_k})\mu_{r-2}(g^{(3)}_{\hat{\tau}_\ell}).$$

Plugging this into (85) yields

$$\sum_{r\geq R} r\left|\max_{k,\ell\in[m]} \mu_r(g_{\tau_k})\mu_r(g^{(1)}_{\hat{\tau}_\ell})\right| \leq \sum_{r\geq R}\frac{1}{\sqrt{r-2}(r-1)}\max_{k,\ell\in[m]}\mu_{r-3}(g^{(3)}_{\tau_k})\mu_{r-2}(g^{(3)}_{\hat{\tau}_\ell})$$
$$\leq \sum_{r\geq R-3}\max_{k\in[m]}\frac{1}{r^{3/2}}\mu_r(g^{(3)}_{\tau_k})^2 + \sum_{r\geq R-2}\max_{\ell\in[m]}\frac{1}{r^{3/2}}\mu_r(g^{(3)}_{\hat{\tau}_\ell})^2,$$

where the second line follows by applying Cauchy-Schwarz. Using Assumption **(M1)**, according to Lemma C.1, then gives $\max_{\tau\in[-\tau_\infty,\tau_\infty]}\mu_r(g^{(3)}_\tau)^2 \leq C$ for all $r \geq 0$ and some constant $C > 0$. Therefore we can conclude with

$$\sum_{r\geq R} r\left|\max_{k,\ell\in[m]}\mu_r(g_{\tau_k})\mu_r(g^{(1)}_{\hat{\tau}_\ell})\right| \leq 2C\sum_{r\geq 1}\frac{1}{r^{3/2}} \leq 6C < \infty. \quad \square$$

Lemma C.8. *Consider weights and approximated weights $(w_k)_{k\in[m]}, (\hat{w}_k)_{k\in[m]}$ of unit norm as before that both fulfill (A2) and (i) in Lemma A.5, as well as shifts $(\tau_k)_{k\in[m]}, (\hat{\tau}_k)_{k\in[m]}$ within $[-\tau_\infty, \tau_\infty]$ for some $\tau_\infty < \infty$. Let $S_{r,\ell}$ be defined as in (76), and assume that g fulfills the Assumption (M1). Then, there exists a constant $C > 0$ such that for $R \geq 9$ we have*

$$\sum_{\ell=1}^{m}\left(\sum_{r\geq R} S_{r,\ell}\right)^2 \leq C\sqrt{m}\Delta_{W,F}^2.$$

Proof. We start by applying the Cauchy product to the squared series

$$\begin{aligned}\sum_{\ell=1}^{m}\left(\sum_{r\geq R} S_{r,\ell}\right)^2 &= \sum_{\ell=1}^{m}\left(\sum_{r\geq 0}\sum_{s=0}^{r} S_{r+R-s,\ell}S_{s+R,\ell}\right)\\ &= \sum_{r\geq 0}\sum_{s=0}^{r}\sum_{\ell=1}^{m} S_{r+R-s,\ell}S_{s+R,\ell}\\ &\leq \sqrt{m}\sum_{r\geq 0}\sum_{s=0}^{r}\left(\sum_{\ell=1}^{m} S_{r+R-s,\ell}^2 S_{s+R,\ell}^2\right)^{1/2}.\end{aligned}$$

The inner sum is now controlled by a sequence of inequalities similar to Lemma C.5. Again we denote $\Delta_{k,r} := \mu_r(g_{\tau_k})(\hat{w}_k - w_k)$ and $T_{k,r} := \sum_{i=1}^{r}\left(\hat{w}_k^{\otimes(r-i)}\otimes w_k^{\otimes(i-1)}\right)$, then by applying the same chain of inequality as in the beginning of the proof of Lemma C.5 we receive

$$\begin{aligned}\sum_{\ell=1}^{m} S_{r+R-s,\ell}^2 S_{s+R,\ell}^2 &= \sum_{\ell=1}^{m}\mu_{r+R-s}(g_{\hat{\tau}_\ell}^{(1)})^2\mu_{s+R}(g_{\hat{\tau}_\ell}^{(1)})^2\\ &\quad\cdot\left\langle\sum_{k=1}^{m}\Delta_{k,r+R-s}\otimes T_{k,r+R-s}, \hat{w}_\ell^{\otimes r+R-s}\right\rangle^2\left\langle\sum_{k=1}^{m}\Delta_{k,s+R}\otimes T_{k,s+R}, \hat{w}_\ell^{\otimes s+R}\right\rangle^2\\ &= \sum_{\ell=1}^{m}\mu_{r+R-s}(g_{\hat{\tau}_\ell}^{(1)})^2\mu_{s+R}(g_{\hat{\tau}_\ell}^{(1)})^2\\ &\quad\cdot\left\langle\left(\sum_{k=1}^{m}\Delta_{k,r+R-s}\otimes T_{k,r+R-s}\right)\otimes\left(\sum_{k=1}^{m}\Delta_{k,s+R}\otimes T_{k,s+R}\right), \hat{w}_\ell^{\otimes r+2R}\right\rangle^2.\end{aligned}$$

As before we now invoke the frame like condition described in Lemma E.2 to attain a bound depending on the upper spectrum of the Grammian $(\widehat{G}_{r+2R})_{ij} = \langle\hat{w}_i, \hat{w}_j\rangle^{r+2R}$. More precisely, by using the shorthand

$$\mu'_{r,s} := \max_{\ell\in[m]}\mu_{r+R-s}(g_{\hat{\tau}_\ell}^{(1)})^2\mu_{s+R}(g_{\hat{\tau}_\ell}^{(1)})^2, \tag{86}$$

we then have

$$\sum_{\ell=1}^{m} S_{r+R-s,\ell}^2 S_{s+R,\ell}^2 \leq \mu'_{r,s}\|\widehat{G}_{r+2R}\|\left\|\left(\sum_{k=1}^{m}\Delta_{k,r+R-s}\otimes T_{k,r+R-s}\right)\otimes\left(\sum_{k=1}^{m}\Delta_{k,s+R}\otimes T_{k,s+R}\right)\right\|_F^2 \tag{87}$$

$$\leq \mu'_{r,s}\|\widehat{G}_{r+2R}\|\left\|\sum_{k=1}^{m}\Delta_{k,r+R-s}\otimes T_{k,r+R-s}\right\|_F^2\left\|\sum_{k=1}^{m}\Delta_{k,s+R}\otimes T_{k,s+R}\right\|_F^2. \tag{88}$$

The two Frobenius norms can now be estimated as in Lemma C.5, more precisely (84), where we also use the shorthands $\Delta_{W,F}, \Delta_{W,O}$ defined in (77) - (78) as well as

$$\mu_{r,s} := \max_{k\in[m]}\mu_{r+R-s}(g_{\tau_k})^2\max_{k\in[m]}\mu_{s+R}(g_{\tau_k})^2.$$

This gives for some absolute constant $C > 0$

$$\begin{aligned}&\mu'_{r,s}\|\widehat{G}_{r+2R}\|\left\|\sum_{k=1}^{m}\Delta_{k,r+R-s}\otimes T_{k,r+R-s}\right\|_F^2\left\|\sum_{k=1}^{m}\Delta_{k,s+R}\otimes T_{k,s+R}\right\|_F^2\\ &\leq C\mu'_{r,s}\mu_{r,s}\|\widehat{G}_{r+2R}\|(r+R-s)^2(s+R)^2\\ &\quad\cdot\left(\Delta_{W,F}^2+\left(\frac{\log m}{D}\right)^{(r+R-s-1)/2}\Delta_{W,O}\right)\left(\Delta_{W,F}^2+\left(\frac{\log m}{D}\right)^{(s+R-1)/2}\Delta_{W,O}\right)\end{aligned}$$

$$\leq C\mu'_{r,s}\mu_{r,s}\|\widehat{G}_{r+2R}\|(r+R-s)^2(s+R)^2\left(\Delta^2_{W,F}+\left(\frac{\log m}{D}\right)^{(R-1)/2}\Delta_{W,O}\right)^2.$$

Next we identify the dominant factors and simplify. Due to Lemma E.3 we have for some constant $C>0$ that

$$\|\widehat{G}_{r+2R}\|\leq C\left(1+m\left(\frac{\log m}{D}\right)^{(r+2R)/2}\right)\leq C\left(1+m\left(\frac{\log m}{D}\right)^{9}\right),$$

where the last stop follows since $R\geq 9$ and due to $m(\log m)^2\leq D^2$ this can be further simplified to $\|\widehat{G}_{r+2R}\|\leq C$. Similarly, we have

$$\begin{aligned}\left(\frac{\log m}{D}\right)^{(R-1)/2}\Delta_{W,O}&=\left(\frac{\log m}{D}\right)^{(R-1)/2}\sum_{k\neq k'}^{m}\left|\langle\hat{w}_k-w_k,\hat{w}_{k'}-w'_k\rangle\right|\\&\leq\left(\frac{\log m}{D}\right)^{4}m^2\delta^2_{\max}\leq\delta^2_{\max}\leq\Delta^2_{W,F}\end{aligned}$$

and therefore we get

$$\|\widehat{G}_{r+2R}\|\left(\Delta^2_{W,F}+\left(\frac{\log m}{D}\right)^{(R-1)/2}\Delta_{W,O}\right)^2\leq C\Delta^4_{W,F}$$

for some absolute constant $C>0$. Plugging these into (88) results in

$$\sum_{\ell=1}^{m}S^2_{r+R-s,\ell}S^2_{s+R,\ell}\leq C\mu'_{r,s}\mu_{r,s}(r+R-s)^2(s+R)^2\Delta^4_{W,F}.$$

Hence, we have

$$\begin{aligned}\sum_{\ell=1}^{m}\left(\sum_{r\geq R}S_{r,\ell}\right)^2&\leq\sqrt{m}\sum_{r\geq 0}\sum_{s=0}^{r}\left(\sum_{\ell=1}^{m}S^2_{r+R-s,\ell}S^2_{s+R,\ell}\right)^{1/2}\\&\leq C\Delta^2_{W,F}\sqrt{m}\sum_{r\geq 0}\sum_{s=0}^{r}\sqrt{|\mu'_{r,s}\mu_{r,s}|}(r+R-s)(s+R)\\&\leq C\Delta^2_{W,F}\sqrt{m}\left(\sum_{r\geq R}r|\max_{k,\ell}\mu_r(g_{\tau_k})\mu_r(g^{(1)}_{\hat{\tau}_\ell})|\right)^2.\end{aligned}$$

The result then follows by applying Lemma C.7 onto the series in the last line followed by a unification of the constants. □

Now, we are finally able to formalize the key lemma of this section. Recall that

$$\begin{aligned}\Delta_{W,F}&:=\|\widehat{W}-W\|_F\\\Delta_{W,O}&:=\sum_{k\neq k'}^{m}\left|\langle\hat{w}_k-w_k,\hat{w}_{k'}-w_{k'}\rangle\right|,\\\Delta_{W,S}&:=\left\|\sum_{k=1}^{m}w_k-\hat{w}_k\right\|_2.\end{aligned}$$

Lemma C.9. *Consider a shallow neural network f with unit norm weights described by W, shifts $\tau_1,\dots,\tau_m\in[-\tau_\infty,\tau_\infty]$ stored in τ and an activation function g that adheres to **(M1)** with $D\leq m$. Furthermore, consider J, J_* given by (68), (70) constructed with $N_{train}\geq m$ network evaluations $y_1,\dots,y_{N_{train}}$ of f where $y_i=f(X_i)$ and $X_1,\dots,X_N\sim\mathcal{N}(0,\mathrm{Id}_D)$. Denote by $\hat{f}$ an approximation to f constructed from parameters $\widehat{W}=[\hat{w}_1|\dots|\hat{w}_m],\hat{\tau}$ as described above with $\|\hat{w}_k\|=1$ for all $k\in[m]$. Then, there exists an absolute constant $C>0$ and D_0 such that, for dimension $D\geq D_0$, the difference between the gradients of J and of the idealized objective J_* obeys*

$$\left\|\nabla J(\hat{\tau})-\nabla J_*(\hat{\tau})\right\|_2\leq 2\kappa^2\sqrt{m}\,\|\hat{\tau}-\tau\|_2^2+C\Delta_{W,1}+\left(\frac{m^3\delta^2_{\max}t}{N_{train}}\right)^{1/2}\tag{89}$$

for $t>0$ with probability at least $1-2m^2\exp\left(-\frac{t}{C\kappa^4}\right)$ and where

$$\Delta_{W,1}\leq\frac{m^{1/2}\log(m)^{3/4}}{D^{1/4}}\left[\|\widehat{W}-W\|_F+\frac{\Delta^{1/2}_{W,O}}{D^{1/2}}+\left\|\sum_{k=1}^{m}w_k-\hat{w}_k\right\|_2\right].$$

Proof of Lemma C.9. Recall that

$$J(\hat{\tau}) = \frac{1}{2N_{\text{train}}} \sum_{i=1}^{N_{\text{train}}} \Big(\hat{f}(X_i, \hat{\tau}) - f(X_i, \tau)\Big)^2.$$

By chain rule we compute the gradient of J w.r.t. $\hat{\tau}$ as

$$\nabla J(\hat{\tau}) = \frac{1}{N_{\text{train}}} \sum_{i=1}^{N_{\text{train}}} \big(\hat{f}(X_i, \hat{\tau}) - f(X_i, \tau)\big) \nabla \hat{f}(X_i, \hat{\tau}).$$

Adding $0 = (\hat{f}(X_i, \tau) - \hat{f}(X_i, \tau))\nabla \hat{f}(X_i, \hat{\tau})$ to $J(\hat{\tau})$ and applying the triangle inequality to $\nabla J - \nabla J_*$ allows us to separate the error caused by the weight approximation

$$\begin{aligned}
&\Big\|\nabla J(\hat{\tau}) - \nabla J_*(\hat{\tau})\Big\|_2 \\
&\quad \le \Big\| \frac{1}{N_{\text{train}}} \Big(\sum_{i=1}^{N_{\text{train}}} (\hat{f}(X_i, \hat{\tau}) - \hat{f}(X_i, \tau))\nabla \hat{f}(X_i, \hat{\tau}) - \nabla \hat{f}(X_i, \tau)\nabla \hat{f}(X_i, \tau)^\top (\hat{\tau} - \tau) \Big) \Big\|_2 \qquad (90) \\
&\qquad + \Big\| \frac{1}{N_{\text{train}}} \sum_{i=1}^{N_{\text{train}}} \big(\hat{f}(X_i, \tau) - f(X_i, \tau)\big) \nabla \hat{f}(X_i, \hat{\tau}) \Big\|_2. \qquad (91)
\end{aligned}$$

To bound the first term in (90) denote $h(\lambda) = (1-\lambda)\hat{\tau} + \lambda\tau$, then we have

$$\begin{aligned}
\hat{f}(X_i, \hat{\tau}) - \hat{f}(X_i, \tau) &= \sum_{k=1}^{m} g(\hat{w}_k^\top X_i + \tau_k) - g(\hat{w}_k^\top X_i + \hat{\tau}_k) \\
&= \sum_{k=1}^{m} g(\hat{w}_k^\top X_i + h(1)_k) - g(\hat{w}_k^\top X_i + h(0)_k) \\
&= \sum_{k=1}^{m} \int_{h(0)_k}^{h(1)_k} g^{(1)}(\hat{w}_k^\top X_i + u) du \\
&= \sum_{k=1}^{m} \int_0^1 g^{(1)}(\hat{w}_k^\top X_i + h(\lambda)_k) h'(\lambda)_k d\lambda \\
&= \sum_{k=1}^{m} \int_0^1 g^{(1)}(\hat{w}_k^\top X_i + h(\lambda)_k) d\lambda (\tau_k - \hat{\tau}_k).
\end{aligned}$$

Therefore, we can bound (90) as follows:

$$\begin{aligned}
&\Big\| \frac{1}{N_{\text{train}}} \Big(\sum_{i=1}^{N_{\text{train}}} (\hat{f}(X_i, \hat{\tau}) - \hat{f}(X_i, \tau))\nabla \hat{f}(X_i, \hat{\tau}) - \nabla \hat{f}(X_i, \tau)\nabla \hat{f}(X_i, \tau)^\top (\hat{\tau} - \tau) \Big) \Big\|_2 \\
&\le \Big\| \frac{1}{N_{\text{train}}} \Big(\sum_{i=1}^{N_{\text{train}}} \nabla \hat{f}(X_i, \hat{\tau}) \Big(\int_0^1 \nabla \hat{f}(X_i, h(\lambda)) d\lambda \Big)^\top - \nabla \hat{f}(X_i, \tau)\nabla \hat{f}(X_i, \tau)^\top \Big) \Big\| \, \|(\hat{\tau} - \tau)\|_2 .
\end{aligned}$$

Let us fix $\hat{\tau}, \tau$ for now and write the last line in terms of matrices $\hat{F}, F, F^* \in \mathbb{R}^{N_{\text{train}} \times m}$, where the i-th row of these matrices is given by $\nabla \hat{f}(X_i, \hat{\tau}), \nabla \hat{f}(X_i, \tau)$ and $\int_0^1 \nabla \hat{f}(X_i, h(\lambda)) d\lambda$, respectively. We obtain

$$\begin{aligned}
&\Big\| \frac{1}{N_{\text{train}}} \Big(\sum_{i=1}^{N_{\text{train}}} \nabla \hat{f}(X_i, \hat{\tau}) \Big(\int_0^1 \nabla \hat{f}(X_i, h(\lambda)) d\lambda \Big)^\top - \nabla \hat{f}(X_i, \tau)\nabla \hat{f}(X_i, \tau)^\top \Big) \Big\| \, \|(\hat{\tau} - \tau)\|_2 \\
&\le \frac{1}{N_{\text{train}}} \|\hat{F}^\top F^* - F^\top F\| \, \|\hat{\tau} - \tau\|_2 \\
&\le \frac{1}{N_{\text{train}}} \Big(\|\hat{F} - F\| \|F^*\| + \|F\| \|F - F^*\| \Big) \|\hat{\tau} - \tau\|_2 . \qquad (92)
\end{aligned}$$

A simultaneous upper bound for $\|\hat{F} - F\|$ and $\|F - F^*\|$ can be established with elementary matrix arithmetic and the Lipschitz continuity of $g^{(1)}$:

$$\|\hat{F}-F\| \le \|\hat{F}-F\|_F = \Big[\sum_{i=1}^{N_{\text{train}}}\sum_{k=1}^{m}(g^{(1)}(\langle \hat{w}_k, X_i\rangle + \hat{\tau}_k) - g^{(1)}(\langle \hat{w}_k, X_i\rangle + \tau_k))^2\Big]^{\frac{1}{2}}$$
$$\le \left\|g^{(2)}\right\|_\infty \Big[\sum_{i=1}^{N_{\text{train}}}\sum_{k=1}^{m}\left(\hat{\tau}_k-\tau_k\right)^2\Big]^{\frac{1}{2}} = \kappa\sqrt{N_{\text{train}}}\,\|\hat{\tau}-\tau\|_2\,,$$

the same bound follows for $\|F-F^*\|$. A crude bound for $\|F\|$ is given by

$$\|F\| \le \sqrt{N_{\text{train}}m}\max_{ik}|F_{ik}| \le \sqrt{N_{\text{train}}m}\left\|g^{(1)}\right\|_\infty \le \kappa\sqrt{N_{\text{train}}m},$$

the same bound follows for $\|F^*\|$. Hence, we can continue from (92) with

$$\frac{1}{N_{\text{train}}}\Big(\|\hat{F}-F\|\|F^*\| + \|F\|\|F-F^*\|\Big)\|\hat{\tau}-\tau\|_2 \le 2\kappa^2\sqrt{m}\|\hat{\tau}-\tau\|_2^2.$$

The error (91) caused by the difference between $\widehat{W}$ and the original weights W has the form

$$\Big\|\frac{1}{N_{\text{train}}}\sum_{i=1}^{N_{\text{train}}}\left(\hat{f}(X_i,\tau)-f(X_i,\tau)\right)\nabla\hat{f}(X_i,\hat{\tau})\Big\|_2$$
$$=\Big\|\frac{1}{N_{\text{train}}}\sum_{i=1}^{N_{\text{train}}}\Big(\sum_{k=1}^{m}g(X_i^\top\hat{w}_k+\tau_k)-g(X_i^\top w_k+\tau_k)\Big)\nabla\hat{f}(X_i,\hat{\tau})\Big\|_2.$$

Let us define

$$\Delta_W^2 := \Big\|\frac{1}{N_{\text{train}}}\sum_{i=1}^{N_{\text{train}}}\Big(\sum_{k=1}^{m}g(X_i^\top\hat{w}_k+\tau_k)-g(X_i^\top w_k+\tau_k)\Big)\nabla\hat{f}(X_i,\hat{\tau})\Big\|_2^2$$
$$=\sum_{\ell=1}^{m}\left[\frac{1}{N_{\text{train}}}\sum_{i=1}^{N_{\text{train}}}\Big(\sum_{k=1}^{m}g(X_i^\top\hat{w}_k+\tau_k)-g(X_i^\top w_k+\tau_k)\Big)g^{(1)}(\langle X_i,\hat{w}_\ell\rangle+\hat{\tau}_\ell)\right]^2$$

To keep the expressions more compact, we define $Z_{ik\ell} := \varphi_{k,\ell}(X_i)$ and

$$\varphi_{k,\ell}(x) := \Big(g(x^\top\hat{w}_k+\tau_k)-g(x^\top w_k+\tau_k)\Big)g^{(1)}(x^\top\hat{w}_\ell+\hat{\tau}_\ell).$$

Let us also define

$$\Delta_{W,1}^2 := \sum_{\ell=1}^{m}\left[\frac{1}{N_{\text{train}}}\sum_{i=1}^{N_{\text{train}}}\sum_{k=1}^{m}\mathbb{E}[Z_{ik\ell}]\right]^2,$$
$$\Delta_{W,2}^2 := \sum_{\ell=1}^{m}\left[\frac{1}{N_{\text{train}}}\sum_{i=1}^{N_{\text{train}}}\sum_{k=1}^{m}(Z_{ik\ell}-\mathbb{E}[Z_{ik\ell}])\right]^2.$$

Then,

$$\Delta_W^2 = \sum_{\ell=1}^{m}\left[\frac{1}{N_{\text{train}}}\sum_{i=1}^{N_{\text{train}}}\sum_{k=1}^{m}Z_{ik\ell}\right]^2 \le 2\Delta_{W,1}^2 + 2\Delta_{W,2}^2. \tag{93}$$

In what follows we will control $\Delta_{W,1}^2, \Delta_{W,2}^2$ by using Hermite expansions and a concentration argument, respectively.

We begin with $\Delta_{W,2}^2$: The first step is to establish that $Z_{ik\ell}-\mathbb{E}[Z_{ik\ell}]$ is subgaussian and to compute its subgaussian norm. We remark that all expectations for the remainder of this proof are w.r.t. the inputs $X_1,\dots,X_{N_{\text{train}}} \sim \mathcal{N}(0,\mathrm{Id}_D)$. First note that by the mean value theorem there exists values $\xi_{i,k}$ such that

$$Z_{ik\ell} = \langle\hat{w}_k - w_k, X_i\rangle g^{(1)}(x^\top\hat{w}_\ell+\hat{\tau}_\ell)g^{(1)}(\xi_{i,k}),$$

where $g^{(1)}$ is a bounded function according to **(M1)**. We can combine this with the well known property of the sub-Gaussian norm which states that $\|Z_{ik\ell}-\mathbb{E}[Z_{ik\ell}]\|_{\psi_2} \le C\|Z_{ik\ell}\|_{\psi_2}$ for some absolute constant $C>0$. This leads to

$$\|Z_{ik\ell}-\mathbb{E}[Z_{ik\ell}]\|_{\psi_2} \le C\|Z_{ik\ell}\|_{\psi_2} \le C\kappa^2\|\langle\hat{w}_k-w_k,X_i\rangle\|_{\psi_2} \le C\kappa^2\delta_{\max}$$

for all $i\in[N_{\text{train}}]$, $k,\ell\in[m]$ and some absolute constant $C>0$. As a consequence we can apply the general Hoeffding inequality (cf. Theorem 2.6.2 in [52]) which yields the estimate

$$\Delta_{W,2}^2 = \frac{1}{N_{\text{train}}^2}\sum_{\ell=1}^{m}\left(\sum_{k=1}^{m}\sum_{i=1}^{N_{\text{train}}}Z_{ik\ell}-\mathbb{E}[Z_{ik\ell}]\right)^2$$

$$\leq \frac{1}{N_{\text{train}}^2}\sum_{\ell=1}^{m}\left(\sum_{k=1}^{m}\left|\sum_{i=1}^{N_{\text{train}}} Z_{ik\ell}-\mathbb{E}[Z_{ik\ell}]\right|\right)^2 \leq \frac{1}{N_{\text{train}}^2}\sum_{\ell=1}^{m} m^2t^2 = \frac{m^3t^2}{N_{\text{train}}^2},$$

which holds using a union bound with probability at least

$$1-\left(\sum_{k,\ell=1}^{m} 2\exp\left(-\frac{ct^2}{\sum_{i=1}^{N_{\text{train}}}\|Z_{ik\ell}-\mathbb{E}[Z_{ik\ell}]\|_{\psi_2}^2}\right)\right)\geq 1-2m^2\exp\left(-\frac{t^2}{CN_{\text{train}}\delta_{\max}^2\kappa^4}\right),$$

for all $t\geq 0$, where $c, C>0$ are absolute constants. This implies that there exists an absolute constant $C>0$ such that for all $t\geq 0$

$$\mathbb{P}\left(\Delta_{W,2}^2 \leq \frac{m^3\delta_{\max}^2 t}{N_{\text{train}}}\right)\geq 1-2m^2\exp\left(-\frac{t}{C\kappa^4}\right). \tag{94}$$

What remains is to control the means contained in $\Delta_{W,1}^2$. Using the shorthand $g_\tau(\cdot)=g(\cdot+\tau)$ and the Hermite expansion we get

$$\begin{aligned}
\mathbb{E}[Z_{ik\ell}] &= \mathbb{E}\left[(g_{\tau_k}(\hat{w}_k^\top X_i)-g_{\tau_k}(w_k^\top X_i))g_{\hat{\tau}_\ell}^{(1)}(\hat{w}_\ell^\top X_i)\right]\\
&= \mathbb{E}\left[\left(\sum_{r\geq 0}\mu_r(g_{\tau_k})(h_r(\hat{w}_k^\top X_i)-h_r(w_k^\top X_i))\right)\sum_{t\geq 0}\mu_t(g_{\hat{\tau}_\ell}^{(1)})h_t(\hat{w}_\ell^\top X_i)\right]\\
&= \sum_{r\geq 0}\mu_r(g_{\tau_k})\mu_r(g_{\hat{\tau}_\ell}^{(1)})\left(\langle\hat{w}_k,\hat{w}_\ell\rangle^r-\langle w_k,\hat{w}_\ell\rangle^r\right),
\end{aligned}$$

where the last two steps rely on the same properties of the Hermite expansion already used in the previous section. The summand corresponding to $r=0$ in the last line above vanishes, thus we have

$$\Delta_{W,1}^2 = \sum_{\ell=1}^{m}\left[\sum_{k=1}^{m}\sum_{r\geq 1}\mu_r(g_{\tau_k})\mu_r(g_{\hat{\tau}_\ell}^{(1)})\left(\langle\hat{w}_k,\hat{w}_\ell\rangle^r-\langle w_k,\hat{w}_\ell\rangle^r\right)\right]^2.$$

Denote now

$$S_{r,\ell} := \sum_{k=1}^{m}\mu_r(g_{\tau_k})\mu_r(g_{\hat{\tau}_\ell}^{(1)})\left(\langle\hat{w}_k,\hat{w}_\ell\rangle^r-\langle w_k,\hat{w}_\ell\rangle^r\right),$$

then, for any $R\geq 2$, we have

$$\Delta_{W,1}^2 = \sum_{\ell=1}^{m}\left(\sum_{r\geq 1}S_{r,\ell}\right)^2 \leq 2\sum_{\ell=1}^{m}S_{1,\ell}^2+\sum_{r=2}^{R-1}2^r\sum_{\ell=1}^{m}S_{r,\ell}^2+2^R\sum_{\ell=1}^{m}\left(\sum_{r\geq R}S_{r,\ell}\right)^2. \tag{95}$$

Choose now $R=9$ and plug in the result from Lemma C.5, Lemma C.6 and Lemma C.8 which yields for an appropriate constant $C>0$ the bound

$$\Delta_{W,1}^2 \leq Cm\left(\frac{\log m}{D}\right)^{1/2}\left\|\sum_{k=1}^{m}w_k-\hat{w}_k\right\|_2^2+\frac{Cm\log m}{D}\left(\Delta_{W,F}^2+\left(\frac{\log m}{D}\right)^{1/2}\Delta_{W,O}\right) \tag{96}$$

$$+C\sqrt{m}\Delta_{W,F}^2. \tag{97}$$

Reordering the terms and taking the square root we receive

$$\begin{aligned}
\Delta_{W,1} &\leq C\left(m^{1/4}+m^{1/2}\left(\frac{\log m}{D}\right)^{1/2}\right)\Delta_{W,F}+Cm^{1/2}\left(\frac{\log m}{D}\right)^{3/4}\Delta_{W,O}^{1/2}\\
&\qquad +Cm^{1/2}\left(\frac{\log m}{D}\right)^{1/4}\left\|\sum_{k=1}^{m}w_k-\hat{w}_k\right\|_2\\
&\leq C\log(m)^{3/4}\left[\left(m^{1/4}+\frac{m^{1/2}}{D^{1/2}}\right)\Delta_{W,F}+\frac{m^{1/2}}{D^{3/4}}\Delta_{W,O}^{1/2}+\frac{m^{1/2}}{D^{1/4}}\left\|\sum_{k=1}^{m}w_k-\hat{w}_k\right\|_2\right].
\end{aligned}$$

Lastly, we can use

$$m^{1/4}+\frac{m^{1/2}}{D^{1/2}}\leq m^{1/4}+\frac{m^{1/2}}{D^{1/4}}\leq\frac{2m^{1/2}}{D^{1/4}}$$

since $m\geq D$ followed by $\Delta_W\leq C(\Delta_{W,1}+\Delta_{W,2})$ to conclude the proof. Note that we can simply separate the constant that appears in the definition of $\Delta_{W,1}$ to appear outside of $\Delta_{W,1}$, such that we arrive at the formulation appearing in the original statement. □

The previous result shows that the gradients associated with our two objective functions J, J_* fulfill

$$\left\|\nabla J(\hat{\tau}) - \nabla J_*(\hat{\tau})\right\|_2 \leq \kappa^2 \sqrt{m}\, \|\hat{\tau} - \tau\|_2^2 + \Delta_W,$$

according to Lemma C.9, where Δ_W depends on the accuracy of the weight approximation. Next, we leverage this to establish sufficient conditions on the accuracy Δ_W and our initial shift estimate under which both gradient descent iterations will remain close to each other over any number of GD steps. The upcoming proof requires that one of the two gradient descent iterations does converge, which in combination with Lemma C.9 allows to control the other iteration locally. It was already established in Lemma C.4 that A is positive definite in expectation. This suggests that $J_*(\hat{\tau}) = (\hat{\tau} - \tau)^\top A(\hat{\tau} - \tau)$ is strictly convex, provided enough samples N_{train} are used to concentrate A around its expectation E. In particular, strict convexity directly implies that $\hat{\tau}_*^{(n)}$ converges to the true biases τ. We will show this as part of the proof of Theorem 4.3, but for the sake of simplicity we will assume positive definiteness of A in the next statement.

Lemma C.10. *Denote by $\hat{\tau}^{(n)}, \hat{\tau}_*^{(n)}$ the gradient descent iterations given by (69) and (72), respectively. Assume that the objective functions J, J_* defined above fulfill*

$$\|\nabla J(\hat{\tau}) - \nabla J_*(\hat{\tau})\|_2 \leq L\, \|\hat{\tau} - \tau\|_2^2 + \Delta_W, \tag{98}$$

for some $L, \Delta_W \geq 0$ and any $\hat{\tau} \in \mathbb{R}^m$. Furthermore, assume that the matrix A in (71) fulfills $\lambda_{\min} := \lambda_{\min}(A) > 0$. If $\Delta_W \leq \frac{\lambda_{\min}^2}{16L}$ and both gradient descent iterations are started with the same step size $\gamma \leq \|A\|^{-1}$ and from the same initialization $\hat{\tau}^{(0)} = \hat{\tau}_*^{(0)}$, adhering to the bound

$$\|\hat{\tau}^{(0)} - \tau\|_2 \leq \frac{\lambda_{\min}}{4\sqrt{2}L}, \tag{99}$$

then the distance between both iterations at gradient step $n \in \mathbb{N}$ satisfies

$$\|\hat{\tau}^{(n)} - \hat{\tau}_*^{(n)}\|_2 \leq \xi^n \|\hat{\tau}^{(0)} - \tau\|_2 + \frac{2\Delta_W}{\lambda_{\min}}(1 - \xi^n),$$

for $\xi = 1 - \frac{\gamma \lambda_{\min}}{2} \in [0, 1)$.

Proof of Lemma C.10. Plugging in the gradient descent iteration with a simple expansion yields

$$\begin{aligned}
&\left\|\hat{\tau}^{(n+1)} - \hat{\tau}_*^{(n+1)}\right\|_2 \\
&= \left\|\hat{\tau}^{(n)} - \hat{\tau}_*^{(n)} - \gamma\left(\nabla J(\hat{\tau}^{(n)}) - \nabla J_*(\hat{\tau}_*^{(n)})\right)\right\|_2 \\
&= \left\|\hat{\tau}^{(n)} - \hat{\tau}_*^{(n)} - \gamma\left(\nabla J(\hat{\tau}^{(n)}) - \nabla J_*(\hat{\tau}^{(n)})\right) - \gamma\left(\nabla J_*(\hat{\tau}^{(n)}) - \nabla J_*(\hat{\tau}_*^{(n)})\right)\right\|_2 \\
&= \left\|\Big(\mathrm{Id}_m - \gamma A\Big)(\hat{\tau}^{(n)} - \hat{\tau}_*^{(n)}) - \gamma\left(\nabla J(\hat{\tau}^{(n)}) - \nabla J_*(\hat{\tau}^{(n)})\right)\right\|_2 \\
&\leq \left\|\Big(\mathrm{Id}_m - \gamma A\Big)(\hat{\tau}^{(n)} - \hat{\tau}_*^{(n)})\right\|_2 + \gamma\left\|\nabla J(\hat{\tau}^{(n)}) - \nabla J_*(\hat{\tau}^{(n)})\right\|_2,
\end{aligned}$$

where we used the definition of the iterations in the first line followed by a simple expansion and the triangle inequality in the last line. The left term of the last line can be bounded with the spectral norm of $\mathrm{Id}_m - \gamma A$ and the right term according to our initial assumption (98):

$$\begin{aligned}
\|\hat{\tau}^{(n+1)} - \hat{\tau}_*^{(n+1)}\|_2 &\leq \|\mathrm{Id}_m - \gamma A\|\, \|\hat{\tau}^{(n)} - \hat{\tau}_*^{(n)}\|_2 + \gamma L \|\hat{\tau}^{(n)} - \tau\|_2^2 + \gamma \Delta_W \\
&\leq (1 - \gamma \lambda_{\min}) \|\hat{\tau}^{(n)} - \hat{\tau}_*^{(n)}\|_2 + \gamma L \|\hat{\tau}^{(n)} - \tau\|_2^2 + \gamma \Delta_W,
\end{aligned}$$

where the second inequality follows from the bound on the minimal eigenvalue of A. Expanding the right term of the last line with $\hat{\tau}_*^{(n)}$ yields

$$\begin{aligned}
&\|\hat{\tau}^{(n+1)} - \hat{\tau}_*^{(n+1)}\|_2 \\
&\quad \leq (1 - \gamma \lambda_{\min}) \|\hat{\tau}^{(n)} - \hat{\tau}_*^{(n)}\|_2 + \gamma L \|\hat{\tau}^{(n)} - \hat{\tau}_*^{(n)} + \hat{\tau}_*^{(n)} - \tau\|_2^2 + \gamma \Delta_W \\
&\quad \leq (1 - \gamma \lambda_{\min}) \|\hat{\tau}^{(n)} - \hat{\tau}_*^{(n)}\|_2 + 2\gamma L \|\hat{\tau}^{(n)} - \hat{\tau}_*^{(n)}\|_2^2 + 2\gamma L \|\hat{\tau}_*^{(n)} - \tau\|_2^2 + \gamma \Delta_W.
\end{aligned} \tag{100}$$

We can now use the fact that the gradient descent iteration (72) in combination with the convexity of the idealized objective J_* ($\lambda_{\min}(A) > 0$) allows for the recursive bound

$$\begin{aligned}
\|\hat{\tau}_*^{(n)} - \tau\|_2 &= \|\hat{\tau}_*^{(n-1)} - \gamma \nabla J_*(\hat{\tau}_*^{(n-1)}) - \tau\|_2 = \|\hat{\tau}_*^{(n-1)} - \gamma A(\hat{\tau}_*^{(n-1)} - \tau) - \tau\|_2 \\
&= \|(\mathrm{Id}_m - \gamma A)(\hat{\tau}_*^{(n-1)} - \tau)\|_2 \leq \|\mathrm{Id}_m - \gamma A\|\, \|\hat{\tau}_*^{(n-1)} - \tau\|_2
\end{aligned}$$

$$\leq \|\mathrm{Id}_m - \gamma A\|^n \|\hat{\tau}_*^{(0)} - \tau\|_2 \leq (1-\gamma\lambda_{\min})^n \delta_0,$$

where we have denoted by $\delta_0 = \|\hat{\tau}^{(0)} - \tau\|$ the initial error. Plugging this into (100) results in

$$\begin{aligned}\|\hat{\tau}^{(n+1)} - \hat{\tau}_*^{(n+1)}\|_2 \\ \leq (1-\gamma\lambda_{\min})\|\hat{\tau}^{(n)} - \hat{\tau}_*^{(n)}\|_2 + 2\gamma L\|\hat{\tau}^{(n)} - \hat{\tau}_*^{(n)}\|_2^2 + 2\gamma L(1-\gamma\lambda_{\min})^{2n}\delta_0^2 + \gamma\Delta_W.\end{aligned} \tag{101}$$

Define $\Delta_n := \max_{k\leq n} \|\hat{\tau}^{(k)} - \hat{\tau}_*^{(k)}\|_2$. We first show by induction that $\Delta_n \leq \lambda_{\min}/4L$ provided that δ_0 and Δ_W are sufficiently small. For step $n=0$, we have $\|\hat{\tau}^{(0)} - \hat{\tau}_*^{(0)}\|_2 = 0$, so the statement is clearly true. Assume now it holds for n and we have to show the induction step. In other words we have to show $\|\hat{\tau}^{(n+1)} - \hat{\tau}_*^{(n+1)}\|_2 \leq \lambda_{\min}/4L$, so the same bound would hold for Δ_{n+1}. We continue from (101), and get

$$\|\hat{\tau}^{(n+1)} - \hat{\tau}_*^{(n+1)}\|_2 \leq (1-\gamma\lambda_{\min} + 2\gamma L\Delta_n)\|\hat{\tau}^{(n)} - \hat{\tau}_*^{(n)}\|_2 + 2\gamma L(1-\gamma\lambda_{\min})^{2n}\delta_0^2 + \gamma\Delta_W.$$

Using the induction hypothesis $\Delta_n \leq \lambda_{\min}/4L$, this simplifies to

$$\|\hat{\tau}^{(n+1)} - \hat{\tau}_*^{(n+1)}\|_2 \leq \left(1 - \gamma\lambda_{\min}/2\right)\|(\hat{\tau}^{(n)} - \hat{\tau}_*^{(n)})\|_2 + 2\gamma L(1-\gamma\lambda_{\min})^{2n}\delta_0^2 + \gamma\Delta_W.$$

To keep the computation more compact, we will denote

$$\xi := 1 - \frac{\gamma\lambda_{\min}}{2}.$$

Now we can repeat the same computations for $\|\hat{\tau}^{(k)} - \hat{\tau}_*^{(k)}\|_2$, $k \leq n$ as well. This leads to

$$\|\hat{\tau}^{(n+1)} - \hat{\tau}_*^{(n+1)}\|_2 \leq 2\gamma L\delta_0^2 \sum_{k=0}^{n} \xi^k (1-\gamma\lambda_{\min})^{2(n-k)} + \gamma\Delta_W \sum_{k=0}^{n}\xi^k,$$

where we used $\left\|\hat{\tau}^{(0)} - \hat{\tau}_*^{(0)}\right\|_2 = 0$. Both sums are uniformly bounded in n, as can be seen by

$$\begin{aligned}\|\hat{\tau}^{(n+1)} - \hat{\tau}_*^{(n+1)}\|_2 &\leq 2\gamma L\delta_0^2 \frac{\xi^{n+1} - (1-\gamma\lambda_{\min})^{2(n+1)}}{\xi - (1-\gamma\lambda_{\min})^2} + \gamma\Delta_W \frac{1-\xi^{n+1}}{1-\xi} \\ &\leq 2\gamma L\delta_0^2 \frac{\xi^{n+1} - (1-\gamma\lambda_{\min})^{2(n+1)}}{\frac{3}{2}\gamma\lambda_{\min} - \gamma^2\lambda_{\min}^2} + \frac{2\Delta_W}{\lambda_{\min}} \\ &\leq 2L\delta_0^2 \frac{\xi^{n+1}}{\frac{3}{2}\lambda_{\min} - \gamma\lambda_{\min}^2} + \frac{2\Delta_W}{\lambda_{\min}} \leq 4L\delta_0^2 \frac{\xi^{n+1}}{\lambda_{\min}} + \frac{2\Delta_W}{\lambda_{\min}}.\end{aligned} \tag{102}$$

Now we have $4L\delta_0^2\xi^{n+1}\lambda_{\min}^{-1} \leq 4L\delta_0^2\lambda_{\min}^{-1}$. Furthermore, $4L\delta_0^2\lambda_{\min}^{-1} \leq \frac{\lambda_{\min}}{8L}$ as long as

$$\delta_0^2 \leq \frac{\lambda_{\min}^2}{32L^2},$$

which holds according to our initial assumption (99). Similarly, as $\Delta_W \leq \frac{\lambda_{\min}^2}{16L}$ by assumption, we get $\frac{2\Delta_W}{\lambda_{\min}} \leq \frac{\lambda_{\min}}{8L}$ This means we now have

$$\Delta_{n+1} \leq \frac{\lambda_{\min}}{8L} + \frac{\lambda_{\min}}{8L} \leq \frac{\lambda_{\min}}{4L},$$

which concludes the proof of the induction establishing that the two iterations remain close to each other so that $\max_{k\leq n}\|\hat{\tau}^{(k)} - \hat{\tau}_*^{(k)}\|_2 \leq \lambda_{\min}/4L$ for all $n \in \mathbb{N}$. To arrive at the final statement we can continue from (102)

$$\begin{aligned}\|\hat{\tau}^{(n)} - \hat{\tau}_*^{(n)}\|_2 &\leq 2\gamma L\delta_0^2 \frac{\xi^n - (1-\gamma\lambda_{\min})^{2n}}{\xi - (1-\gamma\lambda_{\min})^2} + \gamma\Delta_W \frac{1-\xi^n}{1-\xi} \\ &\leq \frac{4L\delta_0^2}{\lambda_{\min}}\xi^n + \frac{2\Delta_W}{\lambda_{\min}}(1-\xi^n). \quad \square\end{aligned}$$

C.3. Concluding the proof of Theorem 4.3

Theorem 4.3 tells us how accurate the weight approximation and shift initialization has to be such that the initial shifts can be further improved w.h.p. by minimizing the empirical loss $J(\hat{\tau}) = \frac{1}{2N_{\text{train}}}\sum_{i=1}^{N_{\text{train}}}\left(\hat{f}(X_i,\hat{\tau}) - f(X_i,\tau)\right)^2$ on a set of generic inputs via gradient descent. The proof of Theorem 4.3 follows directly by combining Lemma C.4, Lemma C.9 and Lemma C.10. Based on the first result we prove that the idealized gradient descent iteration $\tau_*^{(n)}$ will w.h.p. and linear rate converge to the ground-truth shifts τ

by establishing the strict convexity of J_*. The second set of auxiliary statements (i.e., Lemma C.9-C.10) then shows that the gradient descent iteration derived from the empirical risk $J(\hat{\tau})$ will stay close to $\tau_*^{(n)}$ if weight approximations $\widehat{W}$ and initial shifts $\tau^{(0)}$ are sufficiently accurate.

Proof of Theorem 4.3. Denote $E = \mathbb{E}_{X_1,\dots,X_{N_{\text{train}}}\sim\mathcal{N}(0,\mathrm{Id}_D)}[A]$ with A as in (71), and constructed from inputs $X_1,\dots,X_{N_{\text{train}}} \sim \mathcal{N}(0,\mathrm{Id}_D)$. According to Lemma C.4, there exist constants $\omega, C_1 > 0$, which only depend on g and τ_∞, with

$$\lambda_m(E) \geq \omega - C_1\frac{(m-1)\log^2 m}{D^2} \geq \frac{\omega}{2},$$

provided $(2C_1/\omega)m\log^2 m \leq D^2$, as assumed in Theorem 4.3. Note now that A is a sum of positive semi-definite rank-1 matrices. Thus we can apply the Matrix Chernoff bound in Lemma E.4 to get the concentration bound

$$\mathbb{P}\left(\lambda_m(A) \geq \frac{\lambda_m(E)}{4}\right) \geq 1 - m\cdot 0.7^{\frac{N_{\text{train}}\lambda_m(E)}{R}}, \tag{103}$$

where $R = \sup_{x\in\mathbb{R}^D}\|\nabla\hat{f}(\tau,x)\|_2^2 \leq m\left\|g^{(1)}\right\|_\infty^2 \leq m\kappa^2$. From $0.7 < \exp(-1/3)$ now follows that

$$\mathbb{P}\left(\lambda_m(A) \geq \frac{\omega}{8}\right) \geq 1 - m\cdot\exp\left(-\frac{N_{\text{train}}\omega}{6m\kappa^2}\right). \tag{104}$$

For the remainder of the proof we will condition on the event that the bound in (104) holds.
By the result of Lemma C.9, the difference between the gradients $\nabla J, \nabla J_*$ satisfies

$$\left\|\nabla J(\hat{\tau}) - \nabla J_*(\hat{\tau})\right\|_2 \leq 2\kappa^2\sqrt{m}\,\|\hat{\tau}-\tau\|_2^2 + \Delta_W \tag{105}$$

$$\Delta_W = C\Delta_{W,1} + \left(\frac{m^3\delta_{\max}^2 t}{N_{\text{train}}}\right)^{1/2}, \tag{106}$$

for a constant $C > 0$ and $t > 0$ with probability at least $1 - 2m^2\exp\left(-\frac{t}{C\kappa^4}\right)$ where

$$\Delta_{W,1} \leq \frac{m^{1/2}\log(m)^{3/4}}{D^{1/4}}\left[\|\widehat{W}-W\|_F + \frac{\Delta_{W,O}^{1/2}}{D^{1/2}} + \left\|\sum_{k=1}^m w_k - \hat{w}_k\right\|_2\right].$$

Assuming the event associated with (105) occurs, we can invoke Lemma C.10 with $L = 2\kappa^2\sqrt{m}$ meeting its condition by choosing an appropriate constant C in (24). Then, for a step-size $\gamma \leq 1/\|A\|$, $\lambda_{\min} = \lambda_m(A)$ and $\xi = 1 - \gamma\lambda_{\min}/2$, Lemma C.10 yields

$$\|\hat{\tau}^{(n)} - \hat{\tau}_*^{(n)}\|_2 \leq \xi^n\|\hat{\tau}^{(0)} - \tau\|_2 + C(1-\xi^n)\Delta_W. \tag{107}$$

The bound in (107) controls the deviation of the gradient descent iteration (23) from the idealized gradient descent iteration (72). What remains to be shown is that the idealized iteration converges to the correct parameter τ which follows directly by the lower bound on the minimal eigenvalue $\lambda_{\min}$. In fact, we have $J_*(\hat{\tau}) = (\hat{\tau}-\tau)^\top A(\hat{\tau}-\tau)$ and

$$\begin{aligned}\|\hat{\tau}_*^{(n)} - \tau\|_2 &= \|\hat{\tau}_*^{(n-1)} - \gamma\nabla J_*(\hat{\tau}_*^{(n-1)}) - \tau\|_2 = \|(\mathrm{Id}_D - \gamma A)(\hat{\tau}_*^{(n-1)} - \tau)\|_2\\ &\leq \|\mathrm{Id}_D - \gamma A\|^n\|\hat{\tau}_*^{(0)} - \tau\|_2 \leq (1-\gamma\lambda_{\min})^n\|\hat{\tau}^{(0)} - \tau\|_2.\end{aligned}$$

Applying the triangle inequality to (107) therefore yields

$$\begin{aligned}\|\hat{\tau}^{(n)} - \tau\|_2 &\leq \|\hat{\tau}_*^{(n)} - \tau\|_2 + \|\hat{\tau}^{(n)} - \hat{\tau}_*^{(n)}\|_2\\ &\leq \left(\left(1-\gamma\lambda_{\min}\right)^n + \xi^n\right)\|\tau^{(0)} - \tau\|_2 + C(1-\xi^n)\Delta_W\\ &\leq 2\xi^n\|\hat{\tau}^{(0)} - \tau\|_2 + C(1-\xi^n)\Delta_W.\end{aligned}$$

The main statement follows by a union bound over the events described above and by unifying the involved constants. □

Appendix D. Proof of Theorem 2.1

Proof of Theorem 2.1. According to our assumptions, there exist C, D_0 such that the conditions of Theorem 3.3 are fulfilled, and therefore we conclude that the ground truth weights obey **(A1)-(A3)** of Definition 3.2 and that the weight recovery (Algorithm 2) returns vectors $\mathcal{U}$ such that for all $\hat{w} \in \mathcal{U}$ we have

$$\max_{k\in[m]}\min_{s\in\{-1,+1\}}\left\|\hat{w} - sw_k\right\|_2 \leq C_1(m/\alpha)^{1/4}\epsilon^{1/2}, \tag{108}$$

with probability at least

$$1 - \frac{1}{m} - D^2 \exp\left(-\min\{\alpha, 1\} t / C_1\right) - C_1 \exp(-\sqrt{m}/C_1).$$

Denote the weight approximations obtained in the last step by $\{\hat{w}_1, \dots, \hat{w}_m\} \subset \mathbb{S}^{D-1}$. There exists a permutation π of these vectors such that $w_k \approx \pm \hat{w}_{\pi(k)}$ for all $k \in [m]$. To invoke Proposition 4.2, we now need to make sure that

$$\max_{k\in[m]} \min_{s\in\{-1,+1\}} \left\| \hat{w}_{\pi(k)} - s w_k \right\|_2 \leq \frac{1}{C_2} \frac{D^{1/2}}{m\sqrt{\log m}}. \tag{109}$$

By applying the uniform error bound (108) above, we have

$$C_1 (m/\alpha)^{1/4} \epsilon^{1/2} \leq \frac{1}{C_2} \frac{D^{1/2}}{m\sqrt{\log m}} \Leftrightarrow \epsilon \leq \frac{\sqrt{\alpha}}{C_1^2 C_2} \frac{D}{m^{5/2} \log m},$$

which is guaranteed by our upper bound (9) on ϵ for an appropriate constant. This in turn shows that (109) is met. Hence, by Proposition 4.2, Algorithm 3 returns initial shifts $\hat{\tau}$ such that there exists a $\alpha' \leq \alpha$ such that

$$\begin{aligned} \|\tau - \hat{\tau}\|_2 &\leq C_2 \sqrt{m}\epsilon + C_2 m^{3/2} \left(\frac{\log m}{D}\right)^{3/4} \max_{k\in[m]} \min_{s\in\{-1,+1\}} \left\| \hat{w}_{\pi(k)} - s w_k \right\|_2 \\ &\leq C_2 \sqrt{m}\epsilon + C_2 m^{3/2} \left(\frac{\log m}{D}\right)^{3/4} C_1 (m/\alpha)^{1/4} \epsilon^{1/2} \leq \frac{1}{C m^{1/2}}, \end{aligned}$$

where the last line follows from (9) chosen with an appropriate constant $C > 0$. First, note that this implies that the signs learned by the parameter initialization will be correct. We denote this set of signs as $\bar{s}_1, \dots, \bar{s}_m$. Additionally, the last inequality implies that, for the given step-size, the condition of Theorem 4.3 (see (24)) w.r.t. the error in the initial shift is met. Another criterion that has to be met for Theorem 4.3 is that

$$\frac{C m^{1/2} \log(m)^{3/4}}{D^{1/4}} \left(\|\widehat{W} - W\|_F + \frac{\Delta_{W,O}^{1/2}}{D^{1/2}} + \left\| \sum_{k=1}^m w_k - \hat{w}_k \right\|_2 \right) \leq \frac{1}{C\sqrt{m}}, \tag{110}$$

$$\left(\frac{m^3 \delta_{\max}^2 t}{N_{\text{train}}} \right)^{1/2} \leq \frac{1}{C\sqrt{m}}, \tag{111}$$

where $\Delta_{W,O} = \sum_{k\neq k'}^m \left| \langle w_k - \hat{w}_k, w_{k'} - \hat{w}_{k'} \rangle \right|$. We begin with the upper term and rely on worst case bounds which express the different quantities in terms of the uniform error

$$\delta_{\max} = \max_{k\in[m]} \min_{s\in\{-1,+1\}} \left\| \hat{w}_{\pi(k)} - s w_k \right\|_2,$$

such that

$$\|W - \widehat{W}\|_F \leq m^{1/2} \delta_{\max}, \tag{112}$$

$$\frac{\Delta_{W,O}^{1/2}}{D^{1/2}} \leq \frac{m\delta_{\max}}{D^{1/2}}, \tag{113}$$

$$\left\| \sum_{k=1}^m w_k - \hat{w}_k \right\|_2 \leq m\delta_{\max}. \tag{114}$$

Based on these bounds and after adjusting the constants we can simplify (110) to

$$\delta_{\max} \leq \frac{D^{1/4}}{C m^2 \log(m)^{3/4}} \Leftrightarrow \epsilon \leq \frac{D^{1/2} \alpha^{1/2}}{C m^{9/2} \log(m)^{3/2}},$$

which is covered by our initial assumptions on the accuracy. Note that this implies for (111) by plugging in the bound for $\delta_{\max}$ that

$$\left(\frac{m^3 \delta_{\max}^2 t}{N_{\text{train}}} \right)^{1/2} \leq \left(\frac{t D^{1/2}}{N_{\text{train}} m \log(m)^{3/2}} \right)^{1/2}.$$

Using $N_{\text{train}} \geq m$ and $t = D^{1/2}$ this implies

$$\left(\frac{m^3 \delta_{\max}^2 t}{N_{\text{train}}} \right)^{1/2} \leq \frac{1}{N_{\text{train}}^{1/2} \log(m)^{3/4}} \leq \frac{1}{C m^{1/2}},$$

for D, m sufficiently large. Therefore all conditions of Theorem 4.3 are satisfied. Hence, there exists a constant C_4 such that the gradient descent iteration (69) started from initial shifts $\hat{\tau}^{[0]} = \hat{\tau}$ will produce iterates $\hat{\tau}^{[0]}, \dots, \hat{\tau}^{[N_{\mathrm{GD}}]}$ such that

$$\left\| \tau - \hat{\tau}_{\pi}^{[n]} \right\|_2 \leq \frac{C_4 m^{1/2} \log(m)^{3/4}}{D^{1/4}} \left(\|\widehat{W} - W\|_F + \frac{\Delta_{W,O}^{1/2}}{D^{1/2}} + \left\| \sum_{k=1}^{m} w_k - \hat{w}_k \right\|_2 \right) \tag{115}$$

$$+ \left(\frac{m^3 \delta_{\max}^2 t}{N_{\mathrm{train}}} \right)^{1/2} + C_4 \frac{1}{\sqrt{m}} \xi^n, \tag{116}$$

for all $n \in [N_{\mathrm{GD}}]$, some permutation π and some constant $\xi \in [0, 1)$ with probability at least

$$1 - m \exp(-N_{\mathrm{train}}/C_4 m) - 2m^2 \exp\left(-D^{1/2}/C_4\right).$$

After unifying the constants and using the bound on $\delta_{\max}$, the statement of Theorem 2.1 follows. □

Appendix E. Auxiliary results

Lemma E.1. *Let $g \in L_2(\mathbb{R}, w_H)$ be K-times continuously differentiable and assume*

$$\lim_{t \to \infty} g^{(k)}(t) h_r(t) w_H(t) = \lim_{t \to -\infty} g^{(k)}(t) h_r(t) w_H(t) = 0 \tag{117}$$

for all $0 \leq k \leq K$. For any $r \in \mathbb{N} \cup \{0\}$ and $k \in [0, \dots, K]$ we have

$$\mu_r(g^{(n)}) = \sqrt{\binom{n+r}{r} n!} \mu_{r+n}(g).$$

Proof. The Hermite polynomials, weighted by $\exp(-t^2/2)$, satisfy the relation

$$\frac{d}{dt}\left(h_r(t) \exp\left(-\frac{t^2}{2}\right)\right) = \frac{d}{dt}\left(\sqrt{\frac{1}{r!}}(-1)^r \frac{d^r}{dt^r} \exp\left(-\frac{t^2}{2}\right)\right) = \sqrt{\frac{1}{r!}}(-1)^r \frac{d^{r+1}}{dt^{r+1}} \exp\left(-\frac{t^2}{2}\right)$$

$$= -\sqrt{r+1}\sqrt{\frac{1}{(r+1)!}}(-1)^{r+1} \frac{d^{r+1}}{dt^{r+1}} \exp\left(-\frac{t^2}{2}\right) = -\sqrt{r+1} h_{r+1}(t) \exp\left(-\frac{t^2}{2}\right).$$

Therefore, by applying integration by parts, we obtain

$$\mu_r(g^{(n)}) = \int g^{(n)}(t) h_r(t) w_H(t) dt = \left[g^{(n-1)}(t) h_r(t) w_H(t) \right]_{-\infty}^{\infty} - \int g^{(n-1)}(t) \frac{d}{dt}\left(h_r(t) w_H(t) \right) dt$$

$$= 0 + \sqrt{r+1} \int g^{(n-1)}(t) h_{r+1}(t) w_H(t) dt = \sqrt{r+1} \mu_{r+1}(g^{(n-1)}),$$

where the boundary terms vanish due to (117). Applying the same computation n-times, we obtain

$$\mu_r(g^{(n)}) = \sqrt{\prod_{\ell=1}^{n} (r+\ell)} \mu_{r+n}(f) = \sqrt{\frac{(r+n)!}{r!}} \mu_{r+n}(g). \quad \square$$

Lemma E.2. *Let $w_k \in \mathbb{R}^D$ for $k = 1, \dots, m$, and denote by $G_n \in \mathbb{R}^{m \times m}$ the Grammian matrix associated with $(w_k^{\otimes n})_{k \in [m]}$, which is given by $(G_n)_{ij} = \langle w_i, w_j \rangle^n$. Then, for any n-mode tensor $T \in \mathbb{R}^{D \times \cdots \times D}$, we have*

$$\sum_{k=1}^{m} \langle T, w_k^{\otimes n} \rangle^2 \leq \|G_n\| \|T\|_F^2. \tag{118}$$

Proof. First note that we can express the Frobenius inner product as an ordinary inner product over $\mathbb{R}^{D^n}$ with the help of the $\mathrm{vec}(\cdot)$ operator, since $\langle T, w_k^{\otimes n} \rangle = \langle \mathrm{vec}(T), \mathrm{vec}(w_k^{\otimes n}) \rangle$. Let us denote

$$W_n := \left(\mathrm{vec}(w_1^{\otimes n}) \middle| \dots \middle| \mathrm{vec}(w_m^{\otimes n}) \right) \in \mathbb{R}^{D^n \times m}.$$

Then, the following chain of inequalities holds

$$\sum_{k=1}^{m} \langle T, w_k^{\otimes n} \rangle^2 = \sum_{k=1}^{m} \langle \mathrm{vec}(T), \mathrm{vec}(w_k^{\otimes n}) \rangle^2$$

$$= \sum_{k=1}^{m} \operatorname{vec}(T)^\top \operatorname{vec}(w_k^{\otimes n}) \operatorname{vec}(w_k^{\otimes n})^\top \operatorname{vec}(T)$$
$$= \operatorname{vec}(T)^\top W_n W_n^\top \operatorname{vec}(T)$$
$$\leq \|W_n^\top W_n\| \cdot \|\operatorname{vec}(T)\|_2^2 = \|W_n^\top W_n\| \|T\|_F^2.$$

Since $\|W_n^\top W_n\| = \|G_n\|$, this finishes the proof. □

Lemma E.3. *Let $w_k \in \mathbb{S}^{D-1}$ for $k = 1, \dots, m$ be unit vectors, and denote by $G_n \in \mathbb{R}^{m \times m}$ the Grammian matrix associated with $(w_k^{\otimes n})_{k \in [m]}$, which is given by $(G_n)_{ij} = \langle w_i, w_j \rangle^n$. Assume that the vectors $w_1, \dots, w_m$ fulfill **(A2)** of Definition 3.2, then there exists an absolute constant $C > 0$ only depending on c_2 in **(A2)** such that*

$$\|G_n\| \leq C \left(1 + m \left(\frac{\log m}{D}\right)^{n/2}\right). \tag{119}$$

Proof. The result follows directly by Gershgorin circle theorem since the diagonal elements must be 1 and the off-diagonal elements are bounded in absolute value by $c_2 \left(\frac{\log m}{D}\right)^{n/2}$. □

Lemma E.4. *Let $Z \in \mathbb{R}^d$ be a random vector and assume $\|Z\|_2^2 \leq R$ almost surely. For N independent copies $Z_1, \dots, Z_N$ of Z, define the random matrix*

$$G := \sum_{i=1}^{N} Z_i Z_i^\top.$$

Then, we have

$$\mathbb{P}\left(\lambda_m(G) \geq \frac{\lambda_m(\mathbb{E}G)}{4}\right) \geq 1 - m 0.7^{\frac{\lambda_m(\mathbb{E}G)}{R}}.$$

Proof. The result follows directly from the standard matrix Chernoff bound. □

Data availability

No data was used for the research described in the article.

References

[1] F. Albertini, E.D. Sontag, V. Maillot, Uniqueness of weights for neural networks, in: Artificial Neural Networks with Applications in Speech and Vision, 1993, pp. 115–125.
[2] Z. Allen-Zhu, Y. Li, Z. Song, A convergence theory for deep learning via over-parameterization, in: International Conference on Machine Learning (ICML), 2019.
[3] S. Arora, N. Cohen, N. Golowich, W. Hu, A convergence analysis of gradient descent for deep linear neural networks, in: International Conference on Learning Representations (ICLR), 2018.
[4] S. Arora, N. Cohen, W. Hu, Y. Luo, Implicit regularization in deep matrix factorization, in: Neural Information Processing Systems (NeurIPS), 2019.
[5] P. Auer, M. Herbster, M. Warmuth, Exponentially many local minima for single neurons, in: Neural Information Processing Systems (NIPS), 1996.
[6] B. Bah, H. Rauhut, U. Terstiege, M. Westdickenberg, Learning deep linear neural networks: Riemannian gradient flows and convergence to global minimizers, Inf. Inference (Feb. 2021).
[7] R.B. Bapat, V.S. Sunder, On majorization and Schur products, Linear Algebra Appl. 72 (1985) 107–117.
[8] A.L. Blum, R.L. Rivest, Training a 3-node neural network is NP-complete, Neural Netw. 5 (1) (1992) 117–127.
[9] S. Bombari, M.H. Amani, M. Mondelli, Memorization and optimization in deep neural networks with minimum over-parameterization, in: Advances in Neural Information Processing Systems, 2022.
[10] A. Brutzkus, A. Globerson, Globally optimal gradient descent for a convnet with Gaussian inputs, in: International Conference on Machine Learning (ICML), 2017.
[11] S. Bubeck, R. Eldan, Y.T. Lee, D. Mikulincer, Network size and weights size for memorization with two-layers neural networks, in: Neural Information Processing Systems (NeurIPS), 2020.
[12] M.D. Buhmann, A. Pinkus, Identifying linear combinations of ridge functions, Adv. Appl. Math. 22 (1) (1999) 103–118.
[13] C.K. Chui, X. Lin, Approximation by ridge functions and neural networks with one hidden layer, J. Approx. Theory 70 (2) (1992) 131–141.
[14] T.M. Cover, Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition, IEEE Trans. Electron. Comput. 3 (1965) 326–334.
[15] S.S. Du, J.D. Lee, H. Li, L. Wang, X. Zhai, Gradient descent finds global minima of deep neural networks, in: International Conference on Machine Learning (ICML), 2019.
[16] S.S. Du, J.D. Lee, Y. Tian, When is a convolutional filter easy to learn?, in: International Conference on Learning Representations (ICLR), 2018.
[17] S.S. Du, J.D. Lee, Y. Tian, A. Singh, B. Poczos, Gradient descent learns one-hidden-layer CNN: don't be afraid of spurious local minima, in: International Conference on Machine Learning (ICML), 2018.
[18] S.S. Du, X. Zhai, B. Poczos, A. Singh, Gradient descent provably optimizes over-parameterized neural networks, in: International Conference on Learning Representations (ICLR), 2019.
[19] C. Fefferman, Reconstructing a neural net from its output, Rev. Mat. Iberoam. 10 (1994) 507–555.

[20] C. Fiedler, M. Fornasier, T. Klock, M. Rauchensteiner, Stable recovery of entangled weights: towards robust identification of deep neural networks from minimal samples, Appl. Comput. Harmon. Anal. 62 (2023) 123–172.
[21] M. Fornasier, T. Klock, M. Rauchensteiner, Robust and resource-efficient identification of two hidden layer neural networks, Constr. Approx. 55 (1) (2022) 475–536.
[22] M. Fornasier, K. Schnass, J. Vybíral, Learning functions of few arbitrary linear parameters in high dimensions, Found. Comput. Math. 12 (2) (Apr. 2012) 229–262.
[23] M. Fornasier, J. Vybíral, I. Daubechies, Robust and resource efficient identification of shallow neural networks by fewest samples, Inf. Inference 10 (2) (2021) 625–695.
[24] H. Fu, Y. Chi, Y. Liang, Guaranteed recovery of one-hidden-layer neural networks via cross entropy, IEEE Trans. Signal Process. 68 (2020) 3225–3235.
[25] G.-B. Huang, Learning capability and storage capacity of two-hidden-layer feedforward networks, IEEE Trans. Neural Netw. 14 (2) (2003) 274–281.
[26] A. Jacot, F. Gabriel, C. Hongler, Neural tangent kernel: convergence and generalization in neural networks, in: Neural Information Processing Systems (NeurIPS), 2018.
[27] M. Janzamin, H. Sedghi, A. Anandkumar, Beating the perils of non-convexity: guaranteed training of neural networks using tensor methods, arXiv:1506.08473, 2015.
[28] S. Judd, On the complexity of loading shallow neural networks, J. Complex. 4 (3) (1988) 177–192.
[29] J. Kileel, T. Klock, J. Pereira, Landscape analysis of an improved power method for tensor decomposition, in: Neural Information Processing Systems (NeurIPS), 2021.
[30] J. Kileel, J.M. Pereira, Subspace power method for symmetric tensor decomposition and generalized PCA, arXiv:1912.04007, 2019.
[31] Y. Li, Y. Yuan, Convergence analysis of two-layer neural networks with ReLU activation, in: Neural Information Processing Systems (NeurIPS), 2017.
[32] K.-C. Lin, Nonlinear Sampling Theory and Efficient Signal Recovery, PhD thesis, University of Maryland, 2020.
[33] M. Mondelli, A. Montanari, On the connection between learning two-layer neural networks and tensor decomposition, in: International Conference on Artificial Intelligence and Statistics (AISTATS), 2019.
[34] A. Montanari, Y. Zhong, The interpolation phase transition in neural networks: memorization and generalization under lazy training, arXiv:2007.12826, 2020.
[35] E. Moroshko, B.E. Woodworth, S. Gunasekar, J.D. Lee, N. Srebro, D. Soudry, Implicit bias in deep linear classification: initialization scale vs training accuracy, in: Neural Information Processing Systems (NeurIPS), 2020.
[36] B. Neyshabur, R. Tomioka, N. Srebro, In search of the real inductive bias: on the role of implicit regularization in deep learning, in: International Conference on Learning Representations (ICLR), 2015.
[37] Q. Nguyen, On the proof of global convergence of gradient descent for deep relu networks with linear widths, in: International Conference on Machine Learning (ICML), 2021.
[38] Q. Nguyen, M. Mondelli, Global convergence of deep networks with one wide layer followed by pyramidal topology, in: Neural Information Processing Systems (NeurIPS), 2020.
[39] S. Oymak, M. Soltanolkotabi, Toward moderate overparameterization: global convergence guarantees for training shallow neural networks, IEEE J. Sel. Areas Inf. Theory 1 (1) (2020) 84–105.
[40] P.P. Petrushev, Approximation by ridge functions and neural networks, SIAM J. Math. Anal. 30 (1) (1998) 155–189.
[41] A. Pinkus, Approximation theory of the MLP model in neural networks, Acta Numer. 8 (1999) 143–195.
[42] I. Safran, O. Shamir, Spurious local minima are common in two-layer ReLU neural networks, in: International Conference on Machine Learning (ICML), 2018.
[43] H. Sedghi, A. Anandkumar, Provable methods for training neural networks with sparse connectivity, arXiv:1412.2693, 2014.
[44] M. Soltanolkotabi, Learning ReLUs via gradient descent, in: Neural Information Processing Systems (NeurIPS), 2017.
[45] M. Soltanolkotabi, A. Javanmard, J.D. Lee, Theoretical insights into the optimization landscape of over-parameterized shallow neural networks, IEEE Trans. Inf. Theory 65 (2) (2018) 742–769.
[46] C. Song, A. Ramezani-Kebrya, T. Pethick, A. Eftekhari, V. Cevher, Subquadratic overparameterization for shallow neural networks, in: Neural Information Processing Systems (NeurIPS), 2021.
[47] Z. Song, X. Yang, Quadratic suffices for over-parametrization via matrix Chernoff bound, arXiv:1906.03593, 2020.
[48] D. Soudry, E. Hoffer, M.S. Nacson, S. Gunasekar, N. Srebro, The implicit bias of gradient descent on separable data, J. Mach. Learn. Res. 19 (1) (2018) 2822–2878.
[49] H.J. Sussmann, Uniqueness of the weights for minimal feedforward nets with a given input-output map, Neural Netw. 5 (4) (1992) 589–593.
[50] Y. Tian, An analytical formula of population gradient for two-layered relu network and its applications in convergence and critical point analysis, in: International Conference on Machine Learning (ICML), 2017.
[51] J.A. Tropp, User-friendly tail bounds for sums of random matrices, Found. Comput. Math. 12 (4) (Aug. 2012) 389–434.
[52] R. Vershynin, High-Dimensional Probability: An Introduction with Applications in Data Science, Cambridge University Press, Sept. 2018.
[53] R. Vershynin, Memory capacity of neural networks with threshold and rectified linear unit activations, SIAM J. Math. Data Sci. 2 (4) (2020) 1004–1033.
[54] V. Vlačić, H. Bölcskei, Affine symmetries and neural network identifiability, Adv. Math. 376 (2020) 107485.
[55] H. Weyl, Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung), Math. Ann. 71 (4) (Dec. 1912) 441–479.
[56] B. Woodworth, S. Gunasekar, J.D. Lee, E. Moroshko, P. Savarese, I. Golan, D. Soudry, N. Srebro, Kernel and rich regimes in overparametrized models, in: Conference on Learning Theory (COLT), 2020.
[57] X. Wu, S.S. Du, R. Ward, Global convergence of adaptive gradient methods for an over-parameterized neural network, arXiv:1902.07111, 2019.
[58] C. Yun, S. Sra, A. Jadbabaie, Small nonlinearities in activation functions create bad local minima in neural networks, in: International Conference on Learning Representations (ICLR), 2019.
[59] C. Yun, S. Sra, A. Jadbabaie, Small ReLU networks are powerful memorizers: a tight analysis of memorization capacity, in: Neural Information Processing Systems (NeurIPS), 2019.
[60] C. Zhang, S. Bengio, M. Hardt, B. Recht, O. Vinyals, Understanding deep learning requires rethinking generalization, in: International Conference on Learning Representations (ICLR), 2017.
[61] X. Zhang, Y. Yu, L. Wang, Q. Gu, Learning one-hidden-layer ReLU networks via gradient descent, in: International Conference on Artificial Intelligence and Statistics (AISTATS), 2019.
[62] K. Zhong, Z. Song, P. Jain, P.L. Bartlett, I.S. Dhillon, Recovery guarantees for one-hidden-layer neural networks, in: International Conference on Machine Learning (ICML), 2017.
[63] M. Zhou, R. Ge, C. Jin, A local convergence theory for mildly over-parameterized two-layer neural network, in: Conference on Learning Theory (COLT), 2021.
[64] D. Zou, Y. Cao, D. Zhou, Q. Gu, Gradient descent optimizes over-parameterized deep relu networks, Mach. Learn. 109 (3) (2020) 467–492.
[65] D. Zou, Q. Gu, An improved analysis of training over-parameterized deep neural networks, in: Neural Information Processing Systems (NeurIPS), 2019.