Beyond Linearization: On Quadratic and Higher-Order Approximation of Wide Neural Networks

Anonymous authors
Paper under double-blind review

Abstract

Recent theoretical work has established connections between over-parametrized neural networks and linearized models governed by the Neural Tangent Kernels (NTKs). NTK theory leads to concrete convergence and generalization results, yet the empirical performance of neural networks are observed to exceed their linearized models, suggesting insufficiency of this theory.

Towards closing this gap, we investigate the training of over-parametrized neural networks that are beyond the NTK regime yet still governed by the Taylor expansion of the network. We bring forward the idea of *randomizing* the neural networks, which allows them to escape their NTK and couple with quadratic models. We show that the optimization landscape of randomized two-layer networks are nice and amenable to escaping-saddle algorithms. We prove concrete generalization and expressivity results on these randomized networks, which leads to sample complexity bounds (of learning certain simple functions) that match the NTK and can in addition be better by a dimension factor when mild distributional assumptions are present. We demonstrate that our randomization technique can be generalized systematically beyond the quadratic case, by using it to find networks that are coupled with higher-order terms in their Taylor series.

1 Introduction

Deep Learning has made remarkable impact on a variety of artificial intelligence applications such as computer vision, reinforcement learning, and natural language processing. Though immensely successful, theoretical understanding of deep learning lags behind. It is not understood how nonlinear neural networks can be efficiently trained to approximate complex decision boundaries with a relatively few number of training samples.

There has been a recent surge of research on connecting neural networks trained via gradient descent with the neural tangent kernel (NTK) (Jacot et al., 2018; Du et al., 2018a;b; Chizat & Bach, 2018b; Allen-Zhu et al., 2018a; Arora et al., 2019a;b). This line of analysis proceeds by coupling the training dynamics of the nonlinear network with the training dynamics of its linearization in a local neighborhood of the initialization, and then analyzing the expressiveness and generalization of the network via the corresponding properties of its linearized model.

Though powerful, NTK is not yet a completely satisfying theory for explaining the success of deep learning in practice. In theory, the expressive power of the linearized model is roughly the same as, and thus limited to, that of the corresponding random feature space (Allen-Zhu et al., 2018a; Wei et al., 2019) or the Reproducing Kernel Hilbert Space (RKHS) (Bietti & Mairal, 2019). While these spaces can approximate any regular (e.g. bounded Lipschitz) function up to arbitrary accuracy, the norm of the approximators can be exponentially large in the feature dimension for certain non-smooth but very simple functions such as a single ReLU (Yehudai & Shamir, 2019). Using NTK analyses, the sample complexity bound for learning these functions can be poor whereas experimental evidence suggests that the sample complexity is mild (Livni et al., 2014). In practice, kernel machines with the NTK have been experimentally demonstrated to yield competitive results on large-scale tasks such as image classification on CIFAR-10; yet there is still a non-neglible performance gap between NTK and full training on the same convolutional architecture (Arora et al., 2019a; Lee et al., 2019). It is an increasingly compelling question whether we can establish theories for training neural networks beyond the NTK regime.

In this paper, we study the optimization and generalization of over-parametrized two-layer neural networks via relating to their *higher-order approximations*, a principled generalization of the NTK. Our theory starts from the fact that a two-layer neural network $f_{\mathbf{W}_0+\mathbf{W}}(\mathbf{x})$ (with smooth activation) can be Taylor expanded with respect to the weight matrix $\mathbf{W}$ as

$$f_{\mathbf{W}_0+\mathbf{W}}(\mathbf{x}) = \sum_{r=1}^{m} a_r \sigma((\mathbf{w}_{0,r}+\mathbf{w}_r)^\top \mathbf{x}) = \underbrace{\sum_{r=1}^{m} a_r \sigma(\mathbf{w}_{0,r}^\top \mathbf{x})}_{f_{\mathbf{W}_0}(\mathbf{x})} + \sum_{k=1}^{\infty} \underbrace{\sum_{r=1}^{m} a_r \frac{\sigma^{(k)}(\mathbf{w}_{0,r}^\top \mathbf{x})}{k!} (\mathbf{w}_r^\top \mathbf{x})^k}_{f^{(k)}_{\mathbf{W}_0,\mathbf{W}}(\mathbf{x})} .$$

Above, $f_{\mathbf{W}_0}$ does not depend on $\mathbf{W}$, and $f^{(1)}$ corresponds to the NTK model, which is the dominant $\mathbf{W}$-dependent term when $\{\mathbf{w}_r\}$ are small and leads to the coupling between the gradient dynamics for training neural net and its NTK $f^{(1)}$.

Our key observation is that the dominance of $f^{(1)}$ is deduced from comparing the *upper bounds*—rather than the *actual values*—of $f^{(k)}_{\mathbf{W}_0,\mathbf{W}}(\mathbf{x})$. It is a priori possible that there exists a subset of $\mathbf{W}$'s in which the dominating term is not $f^{(1)}$ but some other $f^{(k)}$, $k \geq 2$. If we were able to train in that set, the gradient dynamics would be coupled with the dynamics on $f^{(k)}$ rather than $f^{(1)}$ and thus could be very different. That learning is coupled with $f^{(k)}$ could further offer possibilities for expressing certain functions with parameters of lower complexities, or generalizing better, as $f^{(k)}$ is no longer a linearized model. In this paper, we build on this perspective and identify concrete regimes in which neural net learning is coupled with higher-order $f^{(k)}$'s rather than its linearization.

The contribution of this paper can be summarized as follows.

- We demonstrate that after *randomization*, the linear NTK $f^{(1)}$ is no longer the dominant term, and so the gradient dynamics of the neural net is no longer coupled with NTK. Through a simple sign randomization, the training loss of an over-parametrized two-layer neural network can be coupled with that of a quadratic model (Section 3). We prove that the randomized neural net loss exhibits a nice optimization landscape in that every second-order stationary point has training loss not much higher than the best quadratic model, making it amenable to efficient minimization (Section 4).

- We establish results on the generalization and expressive power of such randomized neural nets (Section 5). These results lead to sample complexity bounds for learning certain simple functions that matches the NTK without distributional assumptions and are advantageous when mild isotropic assumptions on the feature are present. In particular, using randomized networks, the sample complexity bound for learning polynomials (and their linear combination) on (relatively) uniform base distributions is $O(d)$ lower than using NTK.

- We show that the randomization technique can be generalized to find neural nets that are dominated by the k-th order term in their Taylor series ($k > 2$), thereby showing the potential of studying the theory of learning over-parametrized models through their *higher-order NTKs* (Section 6).

1.1 Prior work

We review prior work on the optimization, generalization, and expressivity of neural networks.

Neural Net and Kernel Methods Neal (1996) first proposed the connection between infinite-width networks and kernel methods. Later work (Daniely et al., 2016; Williams, 1997; Lee et al., 2018; Novak et al., 2019; Matthews et al., 2018) extended this connection to various settings including deep networks and deep convolutional networks. These works established that gradient descent on only the output layer weights is well-approximated by a kernel method for large width.

More recently, several groups discovered the connection between gradient descent on all the parameters and the neural tangent kernel (Jacot et al., 2018). Li & Liang (2018); Du et al. (2018b) utilized the coupling of the gradient dynamics to prove that gradient descent finds global minimizers of the training loss of two-layer networks, and Du et al. (2018a); Allen-Zhu et al. (2018b); Zou et al. (2018)

generalized this to deep residual and convolutional networks. Using the NTK coupling, Arora et al. (2019b) proved a generalization error bound that matches the kernel method.

Despite the close theoretical connection between NTK and training deep networks, Arora et al. (2019a); Lee et al. (2019); Chizat & Bach (2018b) empirically found a significant performance gap between NTK and actual training. This gap has been theoretically studied in Wei et al. (2019); Allen-Zhu & Li (2019); Yehudai & Shamir (2019); Ghorbani et al. (2019a) which established that NTK has provably higher generalization error than training the neural net for specific data distributions and architectures.

The idea of randomization is initiated by Allen-Zhu et al. (2018a), who use randomization to provably learn a three-layer network; however it is unclear how the sample complexity of their algorithm compares against the NTK. Inspired by their work, we study the potential gains of coupling with a non-linear approximation over the linear NTK — we compare the performance of a quadratic approximation model with the linear NTK on two-layer networks and find that under mild data assumptions the quadratic approximation reduces sample complexity under mild data assumptions.

Outside the NTK Regime It is believed that the success of SGD is largely due to its algorithmic regularization effects. A large body of work Li et al. (2017); Nacson et al. (2019); Gunasekar et al. (2018b;a; 2017); Woodworth et al. (2019) shows that asymptotically gradient descent converges to a max-margin solution with a strong regularization effect, unlike the NTK regularization[1].

For two-layer networks, a series of works used the mean field method to establish the evolution of the network parameters via a Wasserstein gradient flow (Mei et al., 2018b; Chizat & Bach, 2018a; Wei et al., 2018; Rotskoff & Vanden-Eijnden, 2018; Sirignano & Spiliopoulos, 2018). In the mean field regime, the parameters move significantly from their initialization, unlike NTK regime, however it is unclear if the dynamics converge to solutions of low training loss.

Finally, Li et al. (2019) showed how a combination of large learning rate and injected noise amplifies the regularization from the noise and outperforms the NTK of the corresponding architecture.

Landscape Analysis Many prior works have tried to establish favorable landscape properties such as every local minimum is a global minimum (Ge et al., 2017; Du & Lee, 2018; Soltanolkotabi et al., 2018; Hardt & Ma, 2016; Freeman & Bruna, 2016; Nguyen & Hein, 2017a;b; Haeffele & Vidal, 2015; Venturi et al., 2018). Combining with existing advances in gradient descent avoiding saddle-points (Ge et al., 2015; Lee et al., 2016; Jin et al., 2017), these show that gradient descent find the global minimum. Notably, Du & Lee (2018); Ge et al. (2017) show that gradient descent converges to solutions also of low test error, with lower sample complexity than their corresponding NTKs.

Complexity Bounds Recently, researchers have studied norm-based generalization based (Bartlett et al., 2017; Neyshabur et al., 2015; Golowich et al., 2017), tighter compression-based bounds (Arora et al., 2018), and PAC-Bayes bounds (Dziugaite & Roy, 2017; Neyshabur et al., 2017) that identify properties of the parameter that allow for efficient generalization.

2 Preliminaries

Problem setup We consider the standard supervised learning task, in which we are given a labeled dataset $\mathcal{D} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$, where $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ are sampled i.i.d. from some distribution $\mathbb{P}$, and we wish to find a predictor $f : \mathcal{X} \to \mathcal{Y}$. Without loss of generality, we assume that $\mathcal{X} = \mathbb{S}^{d-1}(B_x) \subset \mathbb{R}^d$ for some $B_x > 0$ (so that the features are d-dimensional with norm B_x.)

Let $\ell : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$ be a loss function such that $\ell(y, 0) \leq 1$, and $z \mapsto \ell(y, z)$ is convex, 1-Lipschitz, and three-times differentiable with the second and third derivatives bounded by one for all $y \in \mathcal{Y}$. This includes for example the logistic and soft hinge loss for classification. We let

$$L(f) := \mathbb{E}_{\mathcal{D}}[\ell(y, f(\mathbf{x}))] := \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(\mathbf{x}_i)) \quad \text{and} \quad L_P(f) := \mathbb{E}_{(\mathbf{x},y)\sim P}[\ell(y, f(\mathbf{x}))]$$

[1] As a concrete example, Woodworth et al. (2019) showed that for matrix completion the NTK solution estimates zero on all unobserved entries and the max-margin solution corresponds to the minimum nuclear norm solution.

denote respectively the empirical risk and population risk for any predictor $f : \mathcal{X} \to \mathcal{Y}$.

Over-parametrized two-layer neural network We consider learning an over-parametrized two-layer neural network of the form

$$f_{\mathbf{W}}(\mathbf{x}) = f_{\mathbf{a},\mathbf{W}}(\mathbf{x}) := \frac{1}{\sqrt{m}}\mathbf{a}^\top \sigma(\mathbf{W}^\top \mathbf{x}) = \frac{1}{\sqrt{m}}\sum_{r=1}^{m} a_r \sigma(\mathbf{w}_r^\top \mathbf{x}), \tag{1}$$

where $\mathbf{W} = [\mathbf{w}_1, \dots, \mathbf{w}_r] \in \mathbb{R}^{d\times m}$ is the first layer and $\mathbf{a} = [a_1, \dots, a_m]^\top \in \mathbb{R}^m$ is the second layer. The $1/\sqrt{m}$ factor is chosen to account for the effect of over-parametrization and is consistent with the NTK-type scaling of (Du et al., 2018b; Arora et al., 2019b). In this paper we fix $\mathbf{a}$ and only train $\mathbf{W}$ (and thus use $f_{\mathbf{W}}$ to denote the network.)

Throughout this paper we assume that the activation is second-order smooth in the following sense.

Assumption A (Smooth activation)**.** *The activation function $\sigma \in C^2(\mathbb{R})$, and there exists some absolute constant $C > 0$ such that $|\sigma'(t)| \le Ct^2$, $|\sigma''(t)| \le C|t|$, and $\sigma''(\cdot)$ is C-Lipschitz.*

An example is the cubic ReLU $\sigma(t) = \mathrm{relu}^3(t) = \max\{t, 0\}^3$. The reason for requiring σ to be higher-order smooth (and thus excluding ReLU) will be made clear in the subsequent text[2].

2.1 Notation

We typically reserve lowercases $a, b, \alpha, \beta, \dots$ for scalars, bold lowercases $\mathbf{a}, \mathbf{b}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \dots$ for vectors, and bold uppercases $\mathbf{A}, \mathbf{B}, \dots$ for matrices. For a matrix $\mathbf{A} = [\mathbf{a}_1, \dots, \mathbf{a}_m] \in \mathbb{R}^{d\times m}$, its $2, p$ norm is defined as $\|\underline{\mathsf{A}}\|_{2,p} := (\sum_{r=1}^m \|\mathbf{a}_r\|_2^p)^{1/p}$ for all $p \in [1, \infty]$. In particular we have $\|\cdot\|_{2,2} = \|\cdot\|_{\mathsf{Fr}}$. We let $\mathsf{B}_{2,p}(R) := \{\mathbf{W} : \|\mathbf{W}\|_{2,p} \le R\}$ denote a $2, p$-norm ball of radius R. We use standard Big-Oh notation: $a = O(b)$ for stating $a \le Cb$ for some absolute constant $C > 0$, and $a = \widetilde{O}(b)$ for $a \le Cb$ where C depends at most logarithmically in b and all other problem parameters. For a twice-differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, $\mathbf{x}_\star$ is called a second-order stationary point if $\nabla f(\mathbf{x}_\star) = \mathbf{0}$ and $\nabla^2 f(\mathbf{x}_\star) \succeq \mathbf{0}$.

3 Escaping NTK via randomization

To motivate our study, we now briefly review the NTK theory for over-parametrized neural nets and provide insights on how to go beyond the NTK regime.

Let $\mathbf{W}_0$ denote the weights in a two-layer neural network at initialization and $\mathbf{W}$ denote its movement from $\mathbf{W}_0$ (so that the current weight matrix is $\mathbf{W}_0 + \mathbf{W}$.) The observation in NTK theory, or the theory of lazy training (Chizat & Bach, 2018b), is that for small $\mathbf{W}$ the neural network $f_{\mathbf{W}_0+\mathbf{W}}$ can be Taylor expanded as

$$\begin{aligned} f_{\mathbf{W}_0+\mathbf{W}}(\mathbf{x}) &= \frac{1}{\sqrt{m}}\sum_{r\le m} a_r \sigma((\mathbf{w}_{0,r} + \mathbf{w}_r)^\top \mathbf{x}) \\ &= \underbrace{\frac{1}{\sqrt{m}}\sum_{r\le m} a_r \sigma(\mathbf{w}_{0,r}^\top \mathbf{x})}_{f_{\mathbf{W}_0}(\mathbf{x})} + \underbrace{\frac{1}{\sqrt{m}}\sum_{r\le m} a_r \sigma'(\mathbf{w}_{0,r}^\top \mathbf{x})(\mathbf{w}_r^\top \mathbf{x})}_{:=f^L_{\mathbf{W}}(\mathbf{x})} + O\left(\frac{1}{\sqrt{m}}\sum_{r\le m} (\mathbf{w}_r^\top \mathbf{x})^2\right), \end{aligned}$$

so that the network can be decomposed as the sum of the initial network $f_{\mathbf{W}_0}$, the linearized model $f^L_{\mathbf{W}}$, and higher order terms. Specifically (ignoring $f_{\mathbf{W}_0}$ for the moment), when m is large and $\|\mathbf{w}_r\|_2 = O(m^{-1/2})$, we expect $f^L_{\mathbf{W}} = O(1)$ and higher order terms to be $o_m(1)$, which is indeed the regime when we train $f_{\mathbf{W}_0+\mathbf{W}}$ via gradient descent. Therefore, the trajectory of training $f_{\mathbf{W}_0+\mathbf{W}}$ is coupled with the trajectory of training $f_{\mathbf{W}_0} + f^L_{\mathbf{W}}$, which is a convex problem and enjoys convergence guarantees (Du et al., 2018b).

[2] We note that the only restrictive requirement in Assumption A is the Lipschitzness of σ'', which guarantees second-order smoothness of the objectives. The bounds on derivatives (and specifically their bound near zero) are merely for technical convenience and can be weakened without hurting the results.

Our goal is to find subsets of $\mathbf{W}$ so that the dominating term is not f^L but something else in the higher order part. The above expansion makes clear that this cannot be achieved through simple fixes such as tuning the leading scale $1/\sqrt{m}$ or the learning rate — the domination of f^L appears to hold so long as the movements $\mathbf{w}_r$ are small.

Randomized coupling with quadratic model We now explain how the idea of *randomization*, initiated in (Allen-Zhu et al., 2018a), can help get rid of the domination of f^L. Let $\mathbf{W}$ be a fixed weight matrix. Suppose for each weight vector $\mathbf{w}_r$, we sample a random variable $\Sigma_{rr} \in \mathbb{R}$ and consider instead the random weight matrix

$$\mathbf{W\Sigma} := \mathbf{W}\mathrm{diag}(\{\Sigma_{rr}\}_{r=1}^m) = [\Sigma_{11}\mathbf{w}_1, \dots, \Sigma_{rr}\mathbf{w}_r],$$

then the second-order Taylor expansion of $f_{\mathbf{W}_0+\mathbf{W\Sigma}}$ can be written as

$$\begin{aligned} f_{\mathbf{W}_0+\mathbf{W\Sigma}}(\mathbf{x}) &= \frac{1}{\sqrt{m}}\sum_{r=1}^m a_r\sigma(\mathbf{w}_{0,r}^\top \mathbf{x}) \\ = f_{\mathbf{W}_0}(\mathbf{x}) &+ \underbrace{\frac{1}{\sqrt{m}}\sum_{r=1}^m a_r\sigma'(\mathbf{w}_{0,r}^\top \mathbf{x})(\Sigma_{rr}\mathbf{w}_r^\top \mathbf{x})}_{=f^L_{\mathbf{W\Sigma}}(\mathbf{x})} + \underbrace{\frac{1}{2\sqrt{m}}\sum_{r=1}^m a_r\sigma''(\mathbf{w}_{0,r}^\top \mathbf{x})\Sigma_{rr}^2(\mathbf{w}_r^\top \mathbf{x})^2}_{:=f^Q_{\mathbf{W\Sigma}}(\mathbf{x})} + \dots, \end{aligned}$$

where we have defined in addition the quadratic part $f^Q_{\mathbf{W\Sigma}}$. Due to the existence of $\{\Sigma_{rr}\}$, each original weight $\mathbf{w}_r$ now gets a weight that is different in f^L and f^Q. Specifically, if we choose

$$\Sigma_{rr} \overset{\mathrm{iid}}{\sim} \mathrm{Unif}\{\pm 1\} \tag{2}$$

to be *random signs*, then we have $\Sigma_{rr}^2 \equiv 1$ and thus $f^Q_{\mathbf{W\Sigma}}(\mathbf{x}) \equiv f^Q_{\mathbf{W}}(\mathbf{x})$, whereas $\mathbb{E}[\Sigma_{rr}] = 0$ so that $\mathbb{E}[f^L_{\mathbf{W\Sigma}}(\mathbf{x})] \equiv 0$. Consequently, f^Q is not affected by such randomization whereas $f^L_{\mathbf{W\Sigma}}$ is now mean zero and thus can have substantially lower magnitude than $f^L_{\mathbf{W}}$.

More precisely, when $\|\mathbf{w}_r\|_2 \asymp m^{-1/4}$, the scalings of f^L and f^Q compares as follows:

- We have $\mathbb{E}_{\mathbf{\Sigma}}[f^L_{\mathbf{W\Sigma}}(\mathbf{x})] = 0$ and

$$\mathbb{E}_{\mathbf{\Sigma}}\left[(f^L_{\mathbf{W\Sigma}}(\mathbf{x}))^2\right] = \frac{1}{m}\sum_{r=1}^m a_r^2\sigma'(\mathbf{w}_{0,r}^\top \mathbf{x})^2(\mathbf{w}_r^\top \mathbf{x})^2 = O\left(\frac{1}{m}\sum_{r=1}^m \|\mathbf{w}_r\|_2^2\right) = O(m^{-1/2}),$$

so we expect $f^L_{\mathbf{W\Sigma}}(\mathbf{x}) = O(m^{-1/4})$ over a random draw of $\mathbf{\Sigma}$.

- The quadratic part scales as

$$f^Q_{\mathbf{W\Sigma}}(\mathbf{x}) = f^Q_{\mathbf{W}}(\mathbf{x}) = \frac{1}{2\sqrt{m}}\sum_{r=1}^m a_r\sigma''(\mathbf{w}_{0,r}^\top \mathbf{x})(\mathbf{w}_r^\top \mathbf{x})^2 = O\left(\frac{1}{\sqrt{m}}\sum_{r=1}^m \|\mathbf{w}_r\|_2^2\right) = O(1).$$

Therefore, at the random weight matrix $\mathbf{W\Sigma}$, f^Q dominates f^L and thus the network is coupled with its quadratic part rather than the linear NTK.

3.1 Learning randomized neural nets

The randomization technique leads to the following recipe for learning $\mathbf{W}$: train $\mathbf{W}$ so that $\|\mathbf{w}_r\|_2 = O(m^{-1/4})$ and $\mathbf{W\Sigma}$ has in expectation low loss. We make this precise by formulating the problem as minimizing a randomized neural net risk.

Randomized risk Let $\widetilde{L} : \mathbb{R}^{d\times m} \to \mathbb{R}$ denote the vanilla empirical risk for learning $f_{\mathbf{W}}$:

$$\widetilde{L}(\mathbf{W}) = \mathbb{E}_{\mathcal{D}}\left[\ell(y, f_{\mathbf{W}_0+\mathbf{W}}(\mathbf{x}))\right],$$

where we have reparametrized the weight matrix into $\mathbf{W}_0 + \mathbf{W}$ so that learning starts at $\mathbf{W} = \mathbf{0}$.

Following our recipe, we now formulate our problem as minimizing the expected risk

$$L(\mathbf{W}) := \mathbb{E}_{\mathbf{\Sigma}}[\widetilde{L}(\mathbf{W}\mathbf{\Sigma})] = \mathbb{E}_{\mathbf{\Sigma},\mathcal{D}}\left[\ell(y, f_{\mathbf{W}_0+\mathbf{W}\mathbf{\Sigma}}(\mathbf{x}))\right],$$

where $\mathbf{\Sigma} \in \mathbb{R}^{m\times m}$ is a diagonal matrix with $\Sigma_{rr} \overset{\text{iid}}{\sim} \text{Unif}\{\pm 1\}$. To encourage $\|\mathbf{w}_r\|_2 = O(m^{-1/4})$ and improve generalization, we consider a regularized version of $\widetilde{L}$ and L with $\ell_{2,4}$ regularization:

$$\widetilde{L}_\lambda(\mathbf{W}) := \widetilde{L}(\mathbf{W}) + \lambda \|\mathbf{W}\|_{2,4}^8 \quad \text{and} \quad L_\lambda(\mathbf{W}) = L(\mathbf{W}) + \lambda \|\mathbf{W}\|_{2,4}^8 = \mathbb{E}_{\mathbf{\Sigma}}[\widetilde{L}_\lambda(\mathbf{W}\mathbf{\Sigma})].$$

We note that the specific norm $\|\cdot\|_{2,4}$ is tied with measuring the average magnitude of $f^Q_{\mathbf{W}}$ and is thus needed, whereas the high power is merely to deal with edge cases and not really essential.

Symmetric initialization We initialize the paramters $(\mathbf{a}, \mathbf{W}_0)$ randomly in the following way: set

$$\begin{aligned} &a_1 = \cdots = a_{m/2} = +1, \;\; a_{m/2+1} = \cdots = a_m = -1, \\ &\mathbf{w}_{0,r} = \mathbf{w}_{0,r+m/2} \overset{\text{iid}}{\sim} \mathsf{N}(0, B_x^{-2} I_d), \; \forall r \in [m/2]^{3}. \end{aligned} \tag{3}$$

Above, we set half of the a_i's as $+1$ and half as -1, and the weights $\mathbf{w}_{0,r}$ are i.i.d. in the $+1$ half and copied exactly into the -1 half. Such an initialization is almost equivalent to i.i.d. random $\mathbf{W}_0$, but has the additional benefit that $f_{\mathbf{W}_0}(\mathbf{x}) \equiv 0$ and also leads to simple expressivity arguments. Our initialization scale B_x^{-2} is chosen so that for a random draw of $\mathbf{w}_0$, we have $\mathbf{w}_0^\top \mathbf{x} \sim \mathsf{N}(0,1)$, which is on average $O(1)$[4]. For technical convenience, we also assume henceforth that the realized $\{\mathbf{w}_{0,r}\}$ satisfies the bound

$$\max_{r\in[m]} (B_x \|\mathbf{w}_{0,r}\|_2) = O\left(\sqrt{d + \log(m/\delta)}\right) = \widetilde{O}(\sqrt{d}). \tag{4}$$

This happens with probability at least $1-\delta$ under random initialization (see proof in Appendix A.3), and ensures that $\max_{r\in[m]} |\mathbf{w}_{0,r}^\top \mathbf{x}| \le \widetilde{O}(\sqrt{d})$ simultaneously for all $\mathbf{x}$.

4 Optimization

In this section, we show that L_λ enjoys a nice optimization landscape.

4.1 Nice landscape of clean risk

As the randomized loss L induces coupling of the neural net $f_{\mathbf{W}_0+\mathbf{W}\mathbf{\Sigma}}$ with the quadratic model $f^Q_{\mathbf{W}}$, we expect its behavior to resemble the behavior of gradient descent on the following *clean risk*:

$$L^Q(\mathbf{W}) := \frac{1}{n}\sum_{i=1}^n \ell(y_i, f^Q_{\mathbf{W}}(\mathbf{x}_i)) = \frac{1}{n}\sum_{i=1}^n \ell\left(y_i, \frac{1}{2\sqrt{m}}\left\langle \mathbf{x}_i\mathbf{x}_i^\top, \mathbf{W}\mathbf{D}_i\mathbf{W}^\top \right\rangle\right).$$

Above, we have defined diagonal matrices $\mathbf{D}_i = \text{diag}(\{a_r \sigma''(\mathbf{w}_{0,r}^\top \mathbf{x}_i)\}_{r\in[m]}) \in \mathbb{R}^{m\times m}$ which are not trained.

We now show that the clean risk L^Q, albeit non-convex, possesses a nice optimization landscape.

Lemma 1 (Landscape of clean risk)**.** *Suppose there exists* $\mathbf{W}_\star \in \mathbb{R}^{d\times m}$ *such that* $L^Q(\mathbf{W}_\star) \le \mathsf{OPT}$*. Let* $\mathbf{\Sigma}' \in \mathbb{R}^{m\times m}$ *be a diagonal matrix with* $\Sigma'_{rr} \overset{\text{iid}}{\sim} \text{Unif}\{\pm 1\}$*, then we have*

$$\begin{aligned} &\mathbb{E}_{\mathbf{\Sigma}'}\left[\nabla^2 L^Q(\mathbf{W})[\mathbf{W}_\star\mathbf{\Sigma}', \mathbf{W}_\star\mathbf{\Sigma}']\right] \\ &\le \left\langle \nabla L^Q(\mathbf{W}), \mathbf{W}\right\rangle - 2(L^Q(\mathbf{W}) - \mathsf{OPT}) + \widetilde{O}\left(dB_x^4 \|\mathbf{W}\|_{2,4}^2 \|\mathbf{W}_\star\|_{2,4}^2 m^{-1}\right). \end{aligned} \tag{5}$$

This result implies that, for $\mathbf{W}$ in a certain ball and large m, every point of higher loss than $\mathbf{W}_\star$ will have either a first-order or a second-order descent direction. In other words, every approximate second-order stationary point of L^Q is also an approximate global minimum. Our proof utilizes the fact that L^Q is similar to the loss function in matrix sensing / learning quadratic neural networks, and builds on recent understandings that the landscapes of these problems are often nice (Soltanolkotabi et al., 2018; Du & Lee, 2018; Allen-Zhu et al., 2018a). The proof is deferred to Appendix B.1.

[4] Our choice covers two commonly used scales in neural net analyses: $B_x = 1$, $\mathbf{w}_{0,r} \sim \mathsf{N}(0, I_d)$ in e.g. (Arora et al., 2019b; Allen-Zhu et al., 2018a); $B_x = \sqrt{d}$, $\mathbf{w}_{0,r} \sim \mathsf{N}(0, I_d/d)$ in e.g. (Ghorbani et al., 2019b).

4.2 Nice landscape of randomized neural net risk

With the coupling between $f_{\mathbf{W}_0+\mathbf{W}\mathbf{\Sigma}}(\mathbf{x})$ and $f^Q_{\mathbf{W}}(\mathbf{x})$ in hand, we expect the risk $L(\mathbf{W}) = \mathbb{E}[\widetilde{L}(\mathbf{W}\mathbf{\Sigma})]$ to enjoy similar guarantees as the clean risk does $L^Q(\mathbf{W})$ in Lemma 1. We make this precise in the following result.

Theorem 2 (Landscape of L)**.** *Suppose there exists $\mathbf{W}_\star \in \mathsf{B}_{2,4}(B_{w,\star})$ such that $L^Q(\mathbf{W}_\star) \le \mathsf{OPT}$, and that*

$$m \ge O\big(\big[B_x^{12}B_w^{12} + d^4B_x^4B_w^4 + d^2B_x^{20}B_w^{20}\big]\epsilon^{-4} + d^5B_x^8B_w^8\epsilon^{-2}\big). \tag{6}$$

for some fixed $\epsilon \in (0,1]$ and $B_w \ge B_{w,\star}$, then for all $\mathbf{W} \in \mathsf{B}_{2,4}(B_w)$, we have

$$\mathbb{E}_{\mathbf{\Sigma}'}\big[\nabla^2 L(\mathbf{W})[\mathbf{W}_\star\mathbf{\Sigma}', \mathbf{W}_\star\mathbf{\Sigma}']\big] \le \langle \nabla L(\mathbf{W}), \mathbf{W}\rangle - 2(L(\mathbf{W}) - \mathsf{OPT}) + \epsilon. \tag{7}$$

As an immediate corollary, we have a similar characterization of the regularized loss L_λ.

Corollary 3 (Landscape of L_λ)**.** *For any $B_w \ge B_{w,\star}$, under the conditions of Theorem 2, we have for all $\lambda > 0$ and all $\mathbf{W} \in \mathsf{B}_{2,4}(B_w)$ that*

$$\begin{aligned}&\mathbb{E}_{\mathbf{\Sigma}'}\big[\nabla^2 L_\lambda(\mathbf{W})[\mathbf{W}_\star\mathbf{\Sigma}', \mathbf{W}_\star\mathbf{\Sigma}']\big]\\ &\le \langle \nabla L_\lambda(\mathbf{W}), \mathbf{W}\rangle - 2(L_\lambda(\mathbf{W}) - \mathsf{OPT}) - \lambda\left\|\mathbf{W}\right\|_{2,4}^8 + C\lambda\left\|\mathbf{W}_\star\right\|_{2,4}^8 + \epsilon,\end{aligned} \tag{8}$$

where $C = O(1)$ is an absolute constant.

Theorem 2 follows directly from Lemma 1 through the coupling between L and L^Q (as well as their gradients and Hessians). Corollary 3 then follows by controlling in addition the effect of the regularizer. The full proof of Theorem 2 and Corollary 3 are deferred to Appendices B.4 and B.5.

We now present our main optimization result, which follows directly from Corollary 3.

Theorem 4 (Optimization of L_λ)**.** *Suppose there exists $\mathbf{W}_\star$ such that*

$$L^Q(\mathbf{W}_\star) \le \mathsf{OPT} \quad \text{and} \quad \left\|\mathbf{W}_\star\right\|_{2,4} \le B_{w,\star} \tag{9}$$

for some $\mathsf{OPT} > 0$. For any $\gamma = \Theta(1)$ and $\epsilon > 0$, we can choose λ suitably and $m \ge \widetilde{O}(\mathrm{poly}(d, B_xB_{w,\star}, \epsilon^{-1}, \gamma^{-1}))$ such that the regularized loss L_λ satisfies the following: any second order stationary point $\widehat{\mathbf{W}}$ has low loss and bounded norm:

$$L_\lambda(\widehat{\mathbf{W}}) \le (1+\gamma)\mathsf{OPT} + \epsilon \quad \text{and} \quad \left\|\widehat{\mathbf{W}}\right\|_{2,4} \le O(B_{w,\star}). \tag{10}$$

Proof sketch. The proof of Theorem 4 consists of two stages: first "localize" any second-order stationary point into a (potentially very big) norm ball using the $\left\|\cdot\right\|_{2,4}^8$ regularizer, then use Corollary 3 in this ball to further deduce that L_λ is low and $\left\|\widehat{\mathbf{W}}\right\|_{2,4} \le O(\left\|\mathbf{W}_\star\right\|_{2,4})$. The full proof is deferred to Appendix B.6.

Efficient optimization through escaping-saddle algorithms Theorem 4 states that when the over-parametrization is enough, any second-order stationary point (SOSP) $\widehat{\mathbf{W}}$ of L_λ has loss competitive with OPT, the performance of best quadratic model. Consequently, algorithms that are able to find SOSPs (escape saddles) such as noisy SGD (Jin et al., 2019) can efficiently minimize L_λ to up to a multiplicative / additive factor of OPT

5 Generalization and Expressivity

We now shift attention to studying the generalization and expressivity of the (randomized) neural net $\widehat{\mathbf{W}}$ learned in Theorem 4.

5.1 Generalization

As $\widehat{\mathbf{W}}$ is always coupled (through randomization) with the quadratic model $f^Q_{\widehat{\mathbf{W}}}$, we begin by studying the generalization of the quadratic model.

Generalization of quadratic models Let

$$\mathcal{F}^Q(B_w) := \left\{\mathbf{x} \mapsto f^Q_{\mathbf{W}}(\mathbf{x}) : \|\mathbf{W}\|_{2,4} \le B_w\right\}$$

denote the class of quadratic models for $\mathbf{W}$ in a $\ell_{2,4}$ ball. We first present a lemma that relates the Rademacher complexity of $\mathcal{F}^Q(B_w)$ to the expected operator norm of certain feature maps.

Lemma 5 (Bounding generalization of f^Q via feature operator norm)**.** *For any non-negative loss ℓ such that $z \mapsto \ell(y,z)$ is 1-Lipschitzz and $\ell(y,0) \le 1$ for all $y \in \mathcal{Y}$, we have the Rademacher complexity bound*

$$\mathbb{E}_{\boldsymbol{\sigma},\mathbf{x}}\left[\sup_{\|\mathbf{W}\|_{2,4}\le B_w}\frac{1}{n}\sum_{i=1}^n \sigma_i \ell(y_i, f^Q_{\mathbf{W}}(\mathbf{x}_i))\right] \le 2B_w^2 \mathbb{E}_{\boldsymbol{\sigma},\mathbf{x}}\left[\max_{r\in[m]}\left\|\frac{1}{n}\sum_{i=1}^n \sigma_i\sigma''(\mathbf{w}_{0,r}^\top\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^\top\right\|_{\rm op}\right] + \frac{1}{\sqrt{n}},$$

where $\sigma_i \overset{\rm iid}{\sim} {\rm Unif}\{\pm 1\}$ are Rademacher variables.

Operator norm based generalization Lemma 5 suggests a possibility for the quadratic model to generalize better than the NTK model: the Rademacher complexity of $\mathcal{F}^Q(B_w)$ depends on the "feature maps" $\frac{1}{n}\sum_{i=1}^n \sigma_i\sigma''(\mathbf{w}_{0,r}^\top\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^\top$ through their matrix operator norm. Compared with the (naive) Frobenius norm based generalization bounds, the operator norm is never worse and can be better when additional structure on $\mathbf{x}$ is present. The proof of Lemma 5 is deferred to Appendix C.1.

We now state our main generalization bound on the (randomized) neural net loss L, which concretizes the above insight.

Theorem 6 (Generalization of randomized neural net loss)**.** *For any data-dependent $\widehat{\mathbf{W}}$ such that $\left\|\widehat{\mathbf{W}}\right\|_{2,4} \le B_w$, we have*

$$\mathbb{E}_{\mathbf{W}_0,\mathcal{D}}\left[L(\widehat{\mathbf{W}}) - L_P(\widehat{\mathbf{W}})\right] \le \widetilde{O}\left(\frac{B_x^2B_w^2M_{x,{\rm op}}}{\sqrt{n}} + \frac{1}{\sqrt{n}}\right) + \widetilde{O}\left(B_x^3B_w^3m^{-1/4} + d^2B_x^2B_w^2m^{-1/2}\right),$$

where $M_{x,{\rm op}} := \left(B_x^{-2}\mathbb{E}_{\mathbf{x}}\left[\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top\right\|_{\rm op}\right]\right)^{1/2}$ is the (rescaled) operator norm of the empirical covariance matrix. In particular, $M_{x,{\rm op}} \le 1$ always holds; if in addition $\mathbf{v}^\top\mathbf{x}$ is $K\sqrt{{\rm Var}(\mathbf{v}^\top\mathbf{x})}$ sub-Gaussian for all $\mathbf{v} \in \mathbb{S}^{d-1}(1)$ and $\kappa({\rm Cov}(\mathbf{x})) \le \kappa$, then $M_{x,{\rm op}} \le \kappa/\sqrt{d}$ whenever $n \ge O(K^4d)$.

The generalization bound in Theorem 6 features two desirable properties:

(1) For large m (e.g. $m \gtrsim n^4$), the bound scales at most logarithmically with the width m, therefore allowing learning with small samples and extreme over-parametrization;

(2) The main term $\widetilde{O}(B_x^2B_w^2M_{x,{\rm op}}/\sqrt{n})$ automatically adapts to properties of the feature distribution and can lower the generalization error than the naive bound by at most $O(1/\sqrt{d})$ without requiring us to tune any hyperparameter. Concretely, we have $M_{x,{\rm op}} \le O(1/\sqrt{d})$ when $\mathbf{x}$ has an isotropic distribution such as ${\rm Unif}(\mathbb{S}^{d-1}(B_x))$ or ${\rm Unif}\{\pm B_x/\sqrt{d}\}^d$.

Theorem 6 follows directly from Lemma 5 and a matrix concentration Lemma. The proof is deferred to Appendix C.2.

5.2 Expressivity and Sample Complexity through Quadratic Models

In order to concretize our generalization result, we now study the expressive power of quadratic models through the concrete example of learning high-degree polynomials.

Theorem 7 (Expressivity of f^Q)**.** *Suppose $\{(a_r, \mathbf{w}_{0,r})\}$ are generated according to the symmetric initialization (3), and $f_\star(\mathbf{x}) = \alpha(\boldsymbol{\beta}^\top\mathbf{x})^p$ where $p - 2 \in \{1\} \cup \{2\ell\}_{\ell\ge 0}$. Suppose further that we use $\sigma(t) = \frac{1}{6}{\rm relu}^3(t)$ (so that $\sigma''(t) = {\rm relu}(t)$), then so long as the width is sufficiently large:*

$$m \ge \widetilde{O}\left(ndp^3\alpha^2(B_x\|\boldsymbol{\beta}\|_2)^{2p}\epsilon^{-2}\right),$$

we have with probability at least $1-\delta$ *(over* $\mathbf{W}_0$*) that there exists* $\mathbf{W}_\star \in \mathbb{R}^{d\times m}$ *such that*

$$\left|L^Q(\mathbf{W}_\star) - L(f_\star)\right| \le \epsilon \quad \text{and} \quad \|\mathbf{W}_\star\|_{2,4}^4 \le B_{w,\star}^4 = O\Big(p^3\alpha^2 B_x^{2(p-2)} \|\boldsymbol{\beta}\|_2^{2p}\,\delta^{-1}\Big).$$

The proof of Theorem 7 is based on a reduction from expressing degree p polynomials using quadratic models to expressing degree $p-2$ polynomials using random feature models. The proof can be found in Appendix C.4.

Comparison between quadratic and linearized (NTK) models We now illustrate our results in Theorem 6 and 7 in two concrete examples, in which we compare the sample complexity bounds of the randomized (quadratic) network and the linear NTK when m is sufficiently large.

Learning a single polynomial. Suppose $f_\star(\mathbf{x}) = \alpha(\boldsymbol{\beta}^\top\mathbf{x})^p$ satisfies $L(f_\star) \le \epsilon$, and we wish to find $\widehat{\mathbf{W}}$ with $O(\epsilon)$ test loss. By Theorem 7 we can choose $\mathbf{W}_\star$ such that $L^Q(\mathbf{W}_\star) \le 2\epsilon$, and by Theorem 4 we can find $\widehat{\mathbf{W}}$ such that $L(\widehat{\mathbf{W}}) \le L_\lambda(\widehat{\mathbf{W}}) \le 3\epsilon$ and $\|\widehat{\mathbf{W}}\|_{2,4} = O(B_{w,\star})$. Take $B_x = 1$, and assume $\mathbf{x}$ is sufficiently isotropic so that $M_{x,\rm op} = O(\frac{1}{\sqrt{d}})$, the sample complexity from Theorem 6 is

$$n \ge \widetilde{O}\Big(\frac{B_x^4 B_w^4 M_{x,\rm op}^2}{\epsilon^2}\Big) = \widetilde{O}\Big(\frac{p^3\alpha^2\|\boldsymbol{\beta}\|_2^{2p}}{d\epsilon^2}\Big) := n_Q.$$

In contrast, the sample complexity for linear NTK (Arora et al., 2019b) to reach ϵ test loss is

$$n \ge \widetilde{O}\Big(\frac{p^2\alpha^2\|\boldsymbol{\beta}\|_2^{2p}}{\epsilon^2}\Big) := n_L.$$

We have $n_Q/n_L = \widetilde{O}(p/d)$, a reduction by a dimension factor unless $p \asymp d$. We note that the above comparison is simply comparing upper bounds, since in general the lower bound on the sample complexity of linear NTK is unknown.

Learning a noisy 2-XOR. Wei et al. (2019) established a sample complexity lower bound of linear NTK of $n \ge n_L = \Omega(d^2/\epsilon^2)$ to achieve ϵ generalization error, which allows for a more rigorous comparison against the quadratic regime.

The ground truth function in 2-XOR is $f_\star(\mathbf{x}) = x_1x_2 = ([(\mathbf{e}_1+\mathbf{e}_2)^\top\mathbf{x}]^2 - [(\mathbf{e}_1-\mathbf{e}_2)^\top\mathbf{x}]^2)/4$, where $\mathbf{x} \in \{\pm 1\}^d$, and $f_\star$ attains constant margin on the training distribution constructed in Wei et al. (2019). By Theorem 7 and an additivity argument[5], $f_\star$ can be approximated by $f^Q_{\mathbf{W}_\star}$ with $B_{w,\star} \le O(\|\boldsymbol{\beta}_1\|_2 + \|\boldsymbol{\beta}_2\|_2) = O(1)$. Thus by Theorem 6 the sample complexity for learning noisy 2-XOR through the randomized net $\widehat{\mathbf{W}}$ is

$$n \ge n_Q = \widetilde{O}\left(\frac{B_x^2 B_{w,\star}^2}{\epsilon^2}\right) = \widetilde{O}\Big(\frac{d}{\epsilon^2}\Big).$$

This is $\widetilde{O}(d)$ better than the *sample complexity lower bound* of linear NTK and thus provably better.

6 Higher-order NTKs

In this section, we demonstrate that our idea of randomization for changing the dynamics of learning neural networks can be generalized systematically — through randomization we are able to obtain over-parametrized neural networks in which the k-th order term dominates the Taylor series. Consider a two-layer neural network with $2m$ neurons and symmetric initialization (cf. (3))

$$f_{\mathbf{W}_0+\mathbf{W}}(\mathbf{x}) = \frac{1}{\sqrt{m}}\sum_{r\le m} \sigma((\mathbf{w}_{0,r}+\mathbf{w}_{+,r})^\top\mathbf{x}) - \sigma((\mathbf{w}_{0,r}+\mathbf{w}_{-,r})^\top\mathbf{x}).$$

Assuming σ is analytic on $\mathbb{R}$ (i.e. it equals its Taylor series at any point), we have

$$f_{\mathbf{W}_0+\mathbf{W}}(\mathbf{x}) = \sum_{k=0}^\infty f^{(k)}_{\mathbf{W}_0,\mathbf{W}}(\mathbf{x}),$$

[5] If (f_1, f_2) can be expressed by $(f^Q_{\mathbf{W}^1_\star}, f^Q_{\mathbf{W}^2_\star})$, then $f = f_1 + f_2$ can be expressed by $f^Q_{\mathbf{W}_\star}$ (in expectation) where $\mathbf{W}_\star = \mathbf{W}^1_\star + \mathbf{W}^2_\star\boldsymbol{\Sigma}$.

where we have defined the *k-th order NTK*

$$f^{(k)}_{\mathbf{W}_0,\mathbf{W}}(\mathbf{x}) := \frac{1}{\sqrt{m}} \sum_{r \le m} \frac{1}{k!} \sigma^{(k)}(\mathbf{w}_{0,r}^\top \mathbf{x}) \big((\mathbf{w}_{+,r}^\top \mathbf{x})^k - (\mathbf{w}_{-,r}^\top \mathbf{x})^k \big).$$

Note that $f^{(0)}(\mathbf{x}) \equiv 0$ due to the symmetric initialization, and $f^{(1)}(\mathbf{x})$ is the standard NTK. For an arbitrary $\mathbf{W}$ such that $\|\mathbf{w}_{+,r}\|_2, \|\mathbf{w}_{-,r}\|_2 = o_m(1)$, we expect that $f^{(1)}(\mathbf{x})$ is the dominating term in the expansion.

6.1 Extracting the k-th order term

We now describe an approach to finding $\mathbf{W}$ so that

$$f_{\mathbf{W}_0+\mathbf{W}}(\mathbf{x}) = f^{(k)}_{\mathbf{W}_0,\mathbf{W}}(\mathbf{x}) + o_m(1),$$

that is, the neural net is approximately the k-th order NTK plus an error term that goes to zero as $m \to \infty$, thereby "escaping" the NTK regime. Our approach builds on the following randomization technique: let z_+, z_- be two random variables (distributions) such that

$$\mathbb{E}[z_+^j] = \mathbb{E}[z_-^j] \ \text{ for } j = 0, 1, \dots, k-1 \quad \text{and} \quad \mathbb{E}[z_+^k] = \mathbb{E}[z_-^k] + 1.$$

Set $(\mathbf{w}_{+,r}, \mathbf{w}_{-,r}) = (z_{+,r}\mathbf{w}_{\star,r}, z_{-,r}\mathbf{w}_{\star,r})$, and take $\|\mathbf{w}_{\star,r}\|_2 = O(m^{-1/2k})$, we have

$$f^{(j)}_{\mathbf{W}_0,\mathbf{W}}(\mathbf{x}) = \frac{1}{\sqrt{m}} \sum_{r \le m} \frac{1}{j!} \sigma^{(j)}(\mathbf{w}_{0,r}^\top \mathbf{x}) \underbrace{(z_{+,r}^j - z_{-,r}^j)}_{\text{mean zero}} \underbrace{(\mathbf{w}_{\star,r}^\top \mathbf{x})^j}_{O(m^{-j/2k})} = O_p(m^{-j/2k})$$

for all $j = 1, \dots, k-1$, and

$$\begin{aligned} f^{(k)}_{\mathbf{W}_0,\mathbf{W}}(\mathbf{x}) &= \frac{1}{\sqrt{m}} \sum_{r \le m} \frac{1}{k!} \sigma^{(k)}(\mathbf{w}_{0,r}^\top \mathbf{x}) \underbrace{(z_{+,r}^k - z_{-,r}^k)}_{\text{mean=1}} \underbrace{(\mathbf{w}_{\star,r}^\top \mathbf{x})^k}_{O(m^{-1/2})} \\ &\approx \mathbb{E}_{\mathbf{w}_0}\left[\frac{1}{k!} \sigma^{(k)}(\mathbf{w}_0^\top \mathbf{x}) (m^{1/2k} \mathbf{w}_\star^\top \mathbf{x})^k \right] = O_P(1), \end{aligned}$$

and

$$f^{(k+1)}_{\mathbf{W}_0,\mathbf{W}}(\mathbf{x}) = \frac{1}{\sqrt{m}} \sum_{r \le m} \frac{1}{(k+1)!} \sigma^{(k+1)}(\mathbf{w}_{0,r}^\top \mathbf{x}) \underbrace{(z_{+,r}^{k+1} - z_{-,r}^{k+1})} \underbrace{(\mathbf{w}_{\star,r}^\top \mathbf{x})^{k+1}}_{O(m^{-(k+1)/2k})} = O_P(m^{-1/2k}).$$

Therefore, with high probability, all $f^{(1)}, \dots, f^{(k-1)}$ as well as the remainder term $f - \sum_{j \le k} f^{(j)}$ has order $O(m^{-1/2k})$, and the k-th order NTK $f^{(k)}$ can express an $O(1)$ function.

Existence through concentration One can imagine using concentration arguments on the above randomization to show the existence of some (deterministic) $\mathbf{W}$ at which the neural net is approximately the k-th order NTK: $\mathbb{E}_{\mathbf{x}}[|f_{\mathbf{W}_0+\mathbf{W}}(\mathbf{x}) - f^{(k)}_{\mathbf{W}_0,\mathbf{W}}(\mathbf{x})|] \le \epsilon_m$, where $\epsilon_m \to 0$ as $m \to \infty$. We would like to leave this as future work.

7 Conclusion

In this paper we proposed and studied the optimization and generalization of over-parametrized neural networks through coupling with higher-order terms in their Taylor series. Through coupling with the quadratic model, we showed that the randomized two-layer neural net has a nice optimization landscape (every second-order stationary point has low loss) and is thus amenable to efficient minimization through escape-saddle style algorithms. These networks enjoy the same expressivity and generalization guarantees as linearized models but in addition can generalize better by a dimension factor when distributional assumptions are present. We extended the idea of randomization to show the existence of neural networks whose Taylor series is dominated by the k-th order term.

We believe our work brings in a number of open questions, such as how to better utilize the expressivity of quadratic models, or whether the study of higher-order expansions can lead to a more satisfying theory for explaining the success of full training. We also note that the Taylor series is only one avenue to obtaining accurate approximations of nonlinear neural networks. It would be of interest to design other approximation schemes for neural networks that are coupled with the network in larger regions of parameter space.

REFERENCES

Zeyuan Allen-Zhu and Yuanzhi Li. What can resnet learn efficiently, going beyond kernels? *arXiv preprint arXiv:1905.10337*, 2019.

Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. *arXiv preprint arXiv:1811.04918*, 2018a.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. *arXiv preprint arXiv:1811.03962*, 2018b.

Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. *arXiv preprint arXiv:1802.05296*, 2018.

Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Ruslan Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. *arXiv preprint arXiv:1904.11955*, 2019a.

Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. *arXiv preprint arXiv:1901.08584*, 2019b.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, pp. 6241–6250, 2017.

Alberto Bietti and Julien Mairal. On the inductive bias of neural tangent kernels. *arXiv preprint arXiv:1905.12173*, 2019.

Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in neural information processing systems*, pp. 3040–3050, 2018a.

Lenaic Chizat and Francis Bach. A note on lazy training in supervised differentiable programming. *arXiv preprint arXiv:1812.07956*, 2018b.

Amit Daniely, Roy Frostig, and Yoram Singer. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. In *Advances In Neural Information Processing Systems*, pp. 2253–2261, 2016.

Simon S Du and Jason D Lee. On the power of over-parametrization in neural networks with quadratic activation. *arXiv preprint arXiv:1803.01206*, 2018.

Simon S Du, Jason D Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. *arXiv preprint arXiv:1811.03804*, 2018a.

Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018b.

Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.

C Daniel Freeman and Joan Bruna. Topology and geometry of half-rectified network optimization. *arXiv preprint arXiv:1611.01540*, 2016.

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points — online stochastic gradient for tensor decomposition. In *Proceedings of The 28th Conference on Learning Theory*, pp. 797–842, 2015.

Rong Ge, Jason D Lee, and Tengyu Ma. Learning one-hidden-layer neural networks with landscape design. *arXiv preprint arXiv:1711.00501*, 2017.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Limitations of lazy training of two-layers neural networks. *arXiv preprint arXiv:1906.08899*, 2019a.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *arXiv preprint arXiv:1904.12191*, 2019b.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. *arXiv preprint arXiv:1712.06541*, 2017.

Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, pp. 6151–6159, 2017.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. *arXiv preprint arXiv:1802.08246*, 2018a.

Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems*, pp. 9461–9471, 2018b.

Benjamin D Haeffele and René Vidal. Global optimality in tensor factorization, deep learning, and beyond. *arXiv preprint arXiv:1506.07540*, 2015.

Moritz Hardt and Tengyu Ma. Identity matters in deep learning. *arXiv preprint arXiv:1611.04231*, 2016.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pp. 8571–8580, 2018.

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M. Kakade, and Michael I. Jordan. How to escape saddle points efficiently. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1724–1732, 2017.

Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M Kakade, and Michael I Jordan. Stochastic gradient descent escapes saddle points efficiently. *arXiv preprint arXiv:1902.04811*, 2019.

Jaehoon Lee, Jascha Sohl-dickstein, Jeffrey Pennington, Roman Novak, Sam Schoenholz, and Yasaman Bahri. Deep neural networks as gaussian processes. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=B1EA-M-0Z.

Jaehoon Lee, Lechao Xiao, Samuel S Schoenholz, Yasaman Bahri, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *arXiv preprint arXiv:1902.06720*, 2019.

Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Conference on Learning Theory*, pp. 1246–1257, 2016.

Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. *arXiv preprint arXiv:1808.01204*, 2018.

Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. *arXiv preprint arXiv:1712.09203*, 2017.

Yuanzhi Li, Colin Wei, and Tengyu Ma. Towards explaining the regularization effect of initial large learning rate in training neural networks. *arXiv preprint arXiv:1907.04595*, 2019.

Roi Livni, Shai Shalev-Shwartz, and Ohad Shamir. On the computational efficiency of training neural networks. In *Advances in neural information processing systems*, pp. 855–863, 2014.

Alexander G de G Matthews, Mark Rowland, Jiri Hron, Richard E Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks. *arXiv preprint arXiv:1804.11271*, 2018.

Song Mei, Yu Bai, Andrea Montanari, et al. The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 46(6A):2747–2774, 2018a.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layers neural networks. In *Proceedings of the National Academy of Sciences*, volume 115, pp. E7665–E7671, 2018b.

Ha Quang Minh, Partha Niyogi, and Yuan Yao. Mercer's theorem, feature maps, and smoothing. In *International Conference on Computational Learning Theory*, pp. 154–168. Springer, 2006.

Mor Shpigel Nacson, Suriya Gunasekar, Jason D Lee, Nathan Srebro, and Daniel Soudry. Lexicographic and depth-sensitive margins in homogeneous and non-homogeneous deep models. *arXiv preprint arXiv:1905.07325*, 2019.

Radford M Neal. Priors for infinite networks. In *Bayesian Learning for Neural Networks*, pp. 29–53. Springer, 1996.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Conference on Learning Theory*, pp. 1376–1401, 2015.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nathan Srebro. A PAC-Bayesian approach to spectrally-normalized margin bounds for neural networks. *arXiv preprint arXiv:1707.09564*, 2017.

Quynh Nguyen and Matthias Hein. The loss surface of deep and wide neural networks. *arXiv preprint arXiv:1704.08045*, 2017a.

Quynh Nguyen and Matthias Hein. The loss surface and expressivity of deep convolutional neural networks. *arXiv preprint arXiv:1710.10928*, 2017b.

Roman Novak, Lechao Xiao, Yasaman Bahri, Jaehoon Lee, Greg Yang, Daniel A. Abolafia, Jeffrey Pennington, and Jascha Sohl-dickstein. Bayesian deep convolutional networks with many channels are gaussian processes. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=B1g30j0qF7.

Grant M Rotskoff and Eric Vanden-Eijnden. Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error. *arXiv preprint arXiv:1805.00915*, 2018.

Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks. *arXiv preprint arXiv:1805.01053*, 2018.

Mahdi Soltanolkotabi, Adel Javanmard, and Jason D Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 65(2):742–769, 2018.

Joel A Tropp et al. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

Luca Venturi, Afonso Bandeira, and Joan Bruna. Neural networks with finite intrinsic dimension have no spurious valleys. *arXiv preprint arXiv:1802.06384*, 2018.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Colin Wei, Jason D Lee, Qiang Liu, and Tengyu Ma. On the margin theory of feedforward neural networks. *arXiv preprint arXiv:1810.05369*, 2018.

Colin Wei, D Lee Jason, Qiang Liu, and Tengyu Ma. Regularization matters: Generalization and optimization of neural nets vs their induced kernel. *arXiv preprint arXiv:1810.05369*, 2019.

Christopher KI Williams. Computing with infinite networks. In *Advances in neural information processing systems*, pp. 295–301, 1997.

Blake Woodworth, Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Kernel and deep regimes in overparametrized models. *arXiv preprint arXiv:1906.05827*, 2019.

Gilad Yehudai and Ohad Shamir. On the power and limitations of random features for understanding neural networks. *arXiv preprint arXiv:1904.00687*, 2019.

Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Stochastic gradient descent optimizes over-parameterized deep ReLU networks. *arXiv preprint arXiv:1811.08888*, 2018.

A TECHNICAL TOOLS

A.1 A MATRIX OPERATOR NORM CONCENTRATION BOUND

Lemma 8 (Variant of Theroem 4.6.1, (Tropp et al., 2015)). *Suppose $\{\mathbf{A}_{r,i}\}_{r\in[m],i\in[n]}$ are fixed symmetric $d\times d$ matrices, and $\{\sigma_i\}_{i\in[n]}\overset{\rm iid}{\sim}\mathrm{Unif}\{\pm1\}$ are Rademacher variables. Letting*

$$\mathbf{Y}_r=\sum_{i=1}^n\sigma_i\mathbf{A}_{r,i},$$

then we have

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[\max_{r\in[m]}\|\mathbf{Y}_r\|_{\rm op}\right]\le 4\sqrt{\max_{r\in[m]}v(\mathbf{Y}_r)\log(2md)},$$

where

$$v(\mathbf{Y}):=\left\|\mathbb{E}_{\boldsymbol{\sigma}}[\mathbf{Y}^2]\right\|_{\rm op}.$$

Proof. Applying the high-probability bound in (Tropp et al., 2015, Theorem 4.6.1) and the union bound, we get

$$\mathbb{P}\left(\max_{r\in[m]}\|\mathbf{Y}_r\|_{\rm op}\ge t\right)\le 2d\sum_{r\in[m]}\exp(-t^2/2v(\mathbf{Y}_r))$$
$$\le 2dm\exp(-t^2/2\max_{r\in[m]}v(\mathbf{Y}_r))=\exp\left(-\frac{t^2}{2\max_{r\in[m]}v(\mathbf{Y}_r)}+\log(2dm)\right).$$

Let $V:=\max_{r\in[m]}v(\mathbf{Y}_r)$, we have by integrating the above bound over t that

$$\mathbb{E}\left[\max_{r\in[m]}\|\mathbf{Y}_r\|_{\rm op}\right]\le\int_0^\infty\min\left\{\exp\left(-\frac{t^2}{2V}+\log(2dm)\right),1\right\}dt$$
$$\le\sqrt{4V\log(2dm)}+\int_{\sqrt{4V\log(2dm)}}^\infty\exp(-t^2/2V+\log(2dm))dt$$
$$\le\sqrt{4V\log(2dm)}+\int_{\sqrt{4V\log(2dm)}}^\infty\exp(-t^2/4V)dt$$
$$\le\sqrt{4V\log(2dm)}+\frac{\sqrt{4\pi V}}{2dm}\le 4\sqrt{V\log(2dm)}.$$

□

A.2 EXPRESSING POLYNOMIALS WITH RANDOM FEATURES

Lemma 9. *Let $\sigma(t)=\mathrm{relu}(t)$ and $\mathbf{w}_0\sim\mathsf{N}(0,B_x^{-2}I_d)$ be Gaussian random features. For any $p\in\{1\}\cup\{2\ell\}_{\ell\ge0}$ and $\boldsymbol{\beta}\in\mathbb{R}^d$, there exists a random variable $a=a(\mathbf{w}_0)$ such that*

$$\mathbb{E}_{\mathbf{w}_0}[\sigma(\mathbf{w}_0^\top\mathbf{x})a]=\alpha(\boldsymbol{\beta}^\top\mathbf{x})^p$$

and a satisfies the ℓ_2 norm bound

$$\mathbb{E}_{\mathbf{w}_0}[a^2]\le 2\pi(p\vee1)^3\alpha^2B_x^{2(p-1)}d\,\|\boldsymbol{\beta}\|_2^{2p}.$$

Proof. Consider the ReLU random feature kernel

$$K(\mathbf{x},\mathbf{x}') = \mathbb{E}_{\mathbf{w}_0\sim\mathsf{N}(0,B_x^{-2}I_d)}[\mathrm{relu}(\mathbf{w}_0^\top\mathbf{x})\mathrm{relu}(\mathbf{w}_0^\top\mathbf{x}')],$$

and let $\mathcal{H}_K$ denote the RKHS associated with this kernel. By the equivalence of feature maps (Minh et al., 2006, Proposition 1), for any feature map $\phi : \mathbb{S}^{d-1}(B_x) \mapsto \mathcal{H}$ (where $\mathcal{H}$ is a Hilbert space) that generates K in the sense that

$$K(\mathbf{x},\mathbf{x}') = \langle\phi(\mathbf{x}),\phi(\mathbf{x}')\rangle_{\mathcal{H}},$$

we have for any function f that

$$\|f\|_{\mathcal{H}_K}^2 = \inf_{a\in\mathcal{H}}\Big\{\|a\|_{\mathcal{H}}^2 : f_\star(x) \equiv \langle a,\phi(\mathbf{x})\rangle\Big\}, \tag{11}$$

and the infimum over a is attainable whenever it is finite.

For the ReLU random feature kernel K, let $u := \mathbf{x}^\top\mathbf{x}'/B_x^2$ and $\mathsf{N}_2(\rho)$ denote a bivariate normal distribution with marginals $\mathsf{N}(0,1)$ and correlation $\rho\in[-1,1]$. We have that

$$\begin{aligned}
K(\mathbf{x},\mathbf{x}') &= \mathbb{E}_{\mathbf{w}_0\sim\mathsf{N}(0,B_x^{-2}I_d)}\big[\mathrm{relu}(\mathbf{w}_0^\top\mathbf{x})\mathrm{relu}(\mathbf{w}_0^\top\mathbf{x}')\big] \\
&= \mathbb{E}_{(\mathbb{Z}_1,Z_2)\sim\mathsf{N}_2(u)}[\mathrm{relu}(Z_1)\mathrm{relu}(Z_2)] \\
&= \frac{1}{2\pi}\Big(u(\pi-\arccos u)+\sqrt{1-u^2}\Big) \\
&= \frac{1}{2\pi}\left(1+\frac{\pi}{2}u+\sum_{\ell=1}^\infty\frac{(2\ell-3)!!}{(2\ell-2)!!(2\ell-1)(2\ell)}u^{2\ell}\right) \\
&= \sum_{p\in\{0,1\}\cup\{2\ell\}_{\ell\ge1}} c_p(\mathbf{x}^\top\mathbf{x}')^pB_x^{-2p} \\
&= \sum_{p\in\{0,1\}\cup\{2\ell\}_{\ell\ge1}} \big\langle\sqrt{c_p}B_x^{-p}\mathbf{x}^{\otimes p},\sqrt{c_p}B_x^{-p}(\mathbf{x}')^{\otimes p}\big\rangle,
\end{aligned}$$

where the constants $\{c_p\}$ satisfy

$$c_0 = 1/(2\pi),\quad c_1 = 1/4,\quad c_{2\ell} \ge \frac{1}{2\pi(2\ell-1)^2(2\ell)}\ \text{ for } \ell\ge1,$$

(and thus $c_p \ge (2\pi(p\vee1)^3)^{-1}$ for all p), and $\mathbf{x}^{\otimes k}\in\mathbb{R}^{d^k}$ denote the k-wise tensor product of $\mathbf{x}$. Therefore, if we define feature map

$$\phi(\mathbf{x}) := \big[\sqrt{c_p}B_x^{-p}\mathbf{x}^{\otimes p}\big]_{p\in\{0,1\}\cup\{2\ell\}_{\ell\ge1}},$$

we have $K(\mathbf{x},\mathbf{x}') = \langle\phi(\mathbf{x}),\phi(\mathbf{x}')\rangle$. With this feature map, the function $f_\star(\mathbf{x}) = \alpha(\beta^\top\mathbf{x})^p$ can be represented as

$$f_\star(\mathbf{x}) \equiv \langle c_\star,\phi(\mathbf{x})\rangle \quad\text{where}\quad c_\star = \Big[0,\dots,0,\alpha\cdot c_p^{-1/2}B_x^p\boldsymbol{\beta}^{\otimes p},0,\dots\Big].$$

Thus by the feature map equivalence (11), we have $f_\star\in\mathcal{H}_K$ and

$$\|f\|_{\mathcal{H}_K}^2 \le \|c_\star\|^2 = \alpha^2c_p^{-1}B_x^{2p}\|\boldsymbol{\beta}\|_2^{2p} \le 2\pi(p\vee1)^3\alpha^2B_x^{2p}\|\boldsymbol{\beta}\|_2^{2p}.$$

Now apply the feature map equivalence (11) again with the random feature map

$$\mathbf{x}\mapsto\big\{\mathrm{relu}(\mathbf{w}_0^\top\mathbf{x})\big\}_{\mathbf{w}_0}$$

(which maps into the inner product space of $\mathbf{w}_0\sim\mathsf{N}(0,B_x^{-2}I_d)$), we conclude that there exists $a = a(\mathbf{w}_0)$ such that $f_\star = \mathbb{E}_{\mathbf{w}_0}[\mathrm{relu}(\mathbf{w}_0^\top\mathbf{x})a]$ and

$$\mathbb{E}_{\mathbf{w}_0}[a^2] \le \|f_\star\|_{\mathcal{H}_K}^2 \le 2\pi(p\vee1)^3\alpha^2B_x^{2p}\|\boldsymbol{\beta}\|_2^{2p}.$$

□

A.3 PROOF OF EQUATION (4)

Let $\mathcal{N}$ be an $1/2$-covering of $\mathbb{S}^{d-1}(1)$. We have $|\mathcal{N}| \le 5^d$ and for any vector $\mathbf{w} \in \mathbb{R}^d$ that $\|\mathbf{w}\|_2 \le 2 \sup_{\mathbf{v}\in\mathcal{N}}(\mathbf{v}^\top \mathbf{w})$ (see e.g. (Mei et al., 2018a, Section A).) We thus have

$$\mathbb{P}\left(\max_{r\in[m]} B_x \|\mathbf{w}_{0,r}\|_2 \ge t\right)$$
$$\le \left(\max_{\mathbf{v}\in\mathcal{N}} B_x(\mathbf{v}^\top \mathbf{w}_{0,r}) \ge t/2\right) \le \exp(-t^2/8 + \log|\mathcal{N}| + \log m) \le \exp(-t^2/8 + d\log 5 + \log m).$$

Setting $t = \sqrt{8(d\log 5 + \log(m/\delta))} = O(\sqrt{d + \log(m/\delta)})$ ensures that the above probability does not exceed δ as desired. □

B PROOFS FOR SECTION 4

B.1 PROOF OF LEMMA 1

Computing the gradient of L^Q, we obtain

$$\nabla L^Q(\mathbf{W}) = \frac{2}{n}\sum_{i=1}^n \ell'(y_i, f^Q_{\mathbf{W}}(\mathbf{x}_i))\frac{1}{2\sqrt{m}}\mathbf{x}_i\mathbf{x}_i^\top \mathbf{W}\mathbf{D}_i.$$

Further computing the Hessian gives

$$\begin{aligned}
\nabla^2 L^Q(\mathbf{W})[\mathbf{W}_\star\mathbf{\Sigma}', \mathbf{W}_\star\mathbf{\Sigma}'] &= \frac{2}{n}\sum_{i=1}^n \ell'(y_i, f^Q_{\mathbf{W}}(\mathbf{x}_i)) \cdot \underbrace{\frac{1}{2\sqrt{m}}\left\langle \mathbf{x}_i\mathbf{x}_i^\top, \mathbf{W}_\star\mathbf{\Sigma}'\mathbf{D}_i\mathbf{\Sigma}'\mathbf{W}_\star^\top\right\rangle}_{f^Q_{\mathbf{W}_\star}(\mathbf{x}_i)} \\
&\quad + \frac{4}{n}\sum_{i=1}^n \ell''(y_i, f^Q_{\mathbf{W}}(\mathbf{x}_i)) \cdot \Bigg(\underbrace{\frac{1}{2\sqrt{m}}\left\langle \mathbf{x}_i\mathbf{x}_i^\top, \mathbf{W}\mathbf{D}_i\mathbf{W}_\star^\top\mathbf{\Sigma}'\right\rangle}_{:=\widetilde{y}_i}\Bigg)^2 \\
&= \underbrace{\frac{2}{n}\sum_{i=1}^n \ell'(y_i, f^Q_{\mathbf{W}}(\mathbf{x}_i)) f^Q_{\mathbf{W}_\star}(\mathbf{x}_i)}_{\mathrm{I}} + \underbrace{\frac{4}{n}\sum_{i=1}^n \ell''(y_i, f^Q_{\mathbf{W}}(\mathbf{x}_i))\widetilde{y}_i^2}_{\mathrm{II}}.
\end{aligned}$$

Taking expectation over $\mathbf{\Sigma}'$, and using that $\ell'' \le 1$, term II can be bounded as

$$\begin{aligned}
&\mathbb{E}_{\mathbf{\Sigma}'}[\mathrm{II}] \le \mathbb{E}_{\mathbf{\Sigma}'}\left[\frac{4}{n}\sum_{i=1}^n \widetilde{y}_i^2\right] \\
&= \mathbb{E}_{\mathbf{\Sigma}',\mathcal{D}}\left[\frac{2}{m}\sum_{r\le m}\sigma''(\mathbf{w}_{0,r}^\top\mathbf{x})^2(\mathbf{w}_r^\top\mathbf{x})^2(\Sigma'_{rr}\mathbf{w}_{\star,r}^\top\mathbf{x})^2\right] \\
&\le C\cdot\mathbb{E}_{\mathcal{D}}\left[\frac{1}{m}\sum_{r\le m}(\mathbf{w}_{0,r}^\top\mathbf{x})^2(\mathbf{w}_r^\top\mathbf{x})^2(\mathbf{w}_{\star,r}^\top\mathbf{x})^2\right] \\
&\le CB_x^4 \max_{r\in[m], i\in[n]}(\mathbf{w}_{0,r}^\top\mathbf{x}_i)^2 \cdot \frac{1}{m}\sum_{r\le m}\|\mathbf{w}_r\|_2^2\|\mathbf{w}_{\star,r}\|_2^2 \\
&\le \widetilde{O}\Big(dB_x^4\|\mathbf{W}\|_{2,4}^2\|\mathbf{W}_\star\|_{2,4}^2 m^{-1}\Big),
\end{aligned}$$

where the last step used Cauchy-Schwarz on $\{\|\mathbf{w}_r\|_2\}$ and $\{\|\mathbf{w}_{\star,r}\|_2\}$.

Term I does not involve $\mathbf{\Sigma}'$ and can be deterministically bounded as

$$\begin{aligned}
\mathrm{I} &= 2\mathbb{E}_{\mathcal{D}}[\ell'(y, f^Q_{\mathbf{W}}(\mathbf{x}))f^Q_{\mathbf{W}_\star}(\mathbf{x})] \\
&= 2\mathbb{E}_{\mathcal{D}}[\ell'(y, f^Q_{\mathbf{W}}(\mathbf{x}))f^Q_{\mathbf{W}}(\mathbf{x})] + 2\mathbb{E}_{\mathcal{D}}[\ell'(y, f^Q_{\mathbf{W}}(\mathbf{x}))(f^Q_{\mathbf{W}_\star}(\mathbf{x}) - f^Q_{\mathbf{W}}(\mathbf{x}))] \\
&\overset{(i)}{\le} \left\langle \nabla L^Q(\mathbf{W}), \mathbf{W} \right\rangle + 2\mathbb{E}_{\mathcal{D}}[\ell(y, f^Q_{\mathbf{W}_\star}(\mathbf{x})) - \ell(y, f^Q_{\mathbf{W}}(\mathbf{x}))] \\
&\overset{(ii)}{=} \left\langle \nabla L^Q(\mathbf{W}), \mathbf{W} \right\rangle - 2(L^Q(\mathbf{W}) - \mathsf{OPT}).
\end{aligned}$$

where (i) follows directly by computing $\left\langle \nabla L^Q(\mathbf{W}), \mathbf{W} \right\rangle$ and the convexity of $z \mapsto \ell(y, z)$, and (ii) follows from the assumption that $L^Q(\mathbf{W}_\star) \le \mathsf{OPT}$. Combining the bounds for terms I and II gives the desired result. □

B.2 Coupling lemmas

Lemma 10 (Bound on f^Q)**.** *For any $\mathbf{W} \in \mathbb{R}^{d\times m}$, the quadratic model $f^Q_{\mathbf{W}}$ satisfies the bound*

$$\left| f^Q_{\mathbf{W}}(\mathbf{x}) \right| \le \widetilde{O}\left(\sqrt{d} B_x^2 \left\| \mathbf{W} \right\|_{2,4}^2 \right)$$

for all $\mathbf{x} \in S^{d-1}(B_x)$.

Proof. We have

$$\begin{aligned}
\left| f^Q_{\mathbf{W}}(\mathbf{x}) \right| &= \left| \frac{1}{\sqrt{m}} \sum_{r\le m} a_r \sigma''(\mathbf{w}_{0,r}^\top \mathbf{x})(\mathbf{w}_r^\top \mathbf{x})^2 \right| \\
&\le \frac{1}{\sqrt{m}} \sum_{r\le m} C\left|\mathbf{w}_{0,r}^\top \mathbf{x}\right| \cdot (\mathbf{w}_r^\top \mathbf{x})^2 \le C\sqrt{m} B_x^2 \max_{r\in[m]} \left|\mathbf{w}_{0,r}^\top \mathbf{x}\right| \cdot \frac{1}{m} \sum_{r\le m} \left\| \mathbf{w}_r \right\|_2^2 \\
&\le C\sqrt{m} B_x^2 \widetilde{O}(\sqrt{d}) \cdot \left(\frac{1}{m} \sum_{r\le m} \left\| \mathbf{w}_r \right\|_2^4 \right)^{1/2} = \widetilde{O}\left(\sqrt{d} B_x^2 R_{w,0} \left\| \mathbf{W} \right\|_{2,4}^2 \right).
\end{aligned}$$

□

Lemma 11 (Coupling between f and f^Q)**.** *We have for all $\mathbf{x} \in \mathbb{S}^{d-1}(B_x)$ that*

(a) $\mathbb{E}_{\mathbf{\Sigma}}[f^L_{\mathbf{W\Sigma}}(\mathbf{x})] = 0$ *and* $\mathbb{E}_{\mathbf{\Sigma}}[(f^L_{\mathbf{W\Sigma}}(\mathbf{x}))^2] \le \widetilde{O}(d^2 B_x^2 \left\| \mathbf{W} \right\|_{2,4}^2 m^{-1/2})$.

(b) $|\Delta^Q_{\mathbf{W\Sigma}}(\mathbf{x})| \le O(B_x^3 \left\| \mathbf{W} \right\|_{2,4}^3 m^{-1/4})$ *(almost surely for all $\mathbf{\Sigma}$.)*

Proof. (a) Recall that

$$f^L_{\mathbf{W\Sigma}}(\mathbf{x}) = \frac{1}{\sqrt{m}} \sum_{r\le m} a_r \sigma'(\mathbf{w}_{0,r}^\top \mathbf{x})(\Sigma_{rr} \mathbf{w}_r^\top \mathbf{x}).$$

As Σ_{rr} has mean zero, we have $\mathbb{E}_{\mathbf{\Sigma}}[f^L] = 0$ and

$$\begin{aligned}
&E_{\mathbf{\Sigma}}[(f^L)^2] = \frac{1}{m} \sum_{r\le m} a_r^2 \sigma'(\mathbf{w}_{0,r}^\top \mathbf{x})^2 (\mathbf{w}_r^\top \mathbf{x})^2 \\
&\overset{(i)}{\le} \frac{1}{m} \sum_{r} C(\mathbf{w}_{0,r}^\top \mathbf{x})^4 (\mathbf{w}_r^\top \mathbf{x})^2 \overset{(ii)}{\le} C \max_{r\in[m]} (\mathbf{w}_{0,r}^\top \mathbf{x})^4 \cdot \frac{1}{m} \sum_{r} B_x^2 \left\| \mathbf{w}_r \right\|_2^2 \\
&\overset{(iii)}{\le} \widetilde{O}(d^2 B_x^2) \cdot \frac{1}{m} \sum_{r\le m} \left\| \mathbf{w}_r \right\|_2^2 \overset{(iv)}{\le} C B_x^6 R_{w,0}^4 \cdot \left(\frac{1}{m} \sum_{r\le m} \left\| \mathbf{w}_r \right\|_2^4 \right)^{1/2} = \widetilde{O}\left(d^2 B_x^2 \left\| \mathbf{W} \right\|_{2,4}^2 m^{-1/2} \right).
\end{aligned}$$

Above, (i) follows from the assumption that $|\sigma'(t)| \le Ct^2$, (ii) is Cauchy-Schwarz, (iii) uses the bound (4), and (iv) uses the power mean inequality on $\left\| \mathbf{w}_r \right\|_2$.

(b) We have by the Lipschitzness of σ'' that

$$\begin{aligned}
|\Delta^Q_{\mathbf{W}}(\mathbf{x})| &= \Big| \frac{1}{\sqrt{m}} \sum_{r\le m} a_r \Big(\sigma((\mathbf{w}_{0,r} + \Sigma_{rr}\mathbf{w}_r)^\top \mathbf{x}) - \sigma(\mathbf{w}_{0,r}^\top \mathbf{x}) \\
&\quad - \sigma'(\mathbf{w}_{0,r}^\top \mathbf{x})(\Sigma_{rr}\mathbf{w}_r^\top \mathbf{x}) - \sigma''(\mathbf{w}_{0,r}^\top \mathbf{x})(\Sigma_{rr}\mathbf{w}_r^\top \mathbf{x})^2 \Big) \Big| \\
&\le \frac{1}{\sqrt{m}} \sum_{r\le m} C|\Sigma_{rr}\mathbf{w}_r^\top \mathbf{x}|^3 \overset{(i)}{\le} C\sqrt{m} B_x^3 \cdot \frac{1}{m} \sum_{r\le m} \|\mathbf{w}_r\|_2^3 \\
&\le C\sqrt{m} B_x^3 \cdot \left(\frac{1}{m} \sum_{r\le m} \|\mathbf{w}_r\|_2^4 \right)^{3/4} \\
&= O\left(B_x^3 \|\mathbf{W}\|_{2,4}^3 m^{-1/4} \right),
\end{aligned}$$

where again (i) uses the power mean inequality on $\|\mathbf{w}_r\|_2$.

□

B.3 Closeness of landscapes

Lemma 12 (L^Q close to L)**.** *We have for all $\mathbf{W} \in \mathbb{R}^{d\times m}$ that*

$$|L(\mathbf{W}) - L^Q(\mathbf{W})| \le \widetilde{O}\Big(B_x^3 \|\mathbf{W}\|_{2,4}^3 m^{-1/4} + d^2 B_x^2 \|\mathbf{W}\|_{2,4}^2 m^{-1/2} \Big).$$

Proof. Recall that

$$L(\mathbf{W}) = \mathbb{E}_{\mathbf{\Sigma},\mathcal{D}}[\ell(y, f_{\mathbf{W}_0 + \mathbf{W}\mathbf{\Sigma}}(\mathbf{x}))] \quad \text{and} \quad L^Q(\mathbf{W}) = \mathbb{E}_{\mathcal{D}}[\ell(y, f^Q_{\mathbf{W}}(\mathbf{x}))].$$

By the 1-Lipschitzness of $z \mapsto \ell(y,z)$ we have

$$\begin{aligned}
\big| L(\mathbf{W}) - L^Q(\mathbf{W}) \big| &\le \mathbb{E}_{\mathbf{\Sigma},\mathcal{D}}\Big[\Big| f_{\mathbf{W}_0 + \mathbf{W}\mathbf{\Sigma}}(\mathbf{x}) - f^Q_{\mathbf{W}}(\mathbf{x}) \Big| \Big] \\
&\le \Big(\mathbb{E}_{\mathbf{\Sigma},\mathcal{D}}\Big[(f^L_{\mathbf{W}\mathbf{\Sigma}}(\mathbf{x}) + \Delta^Q_{\mathbf{W}\mathbf{\Sigma}}(\mathbf{x}))^2 \Big] \Big)^{1/2} \\
&\le \Big(2\mathbb{E}_{\mathbf{\Sigma},\mathcal{D}}\big[(f^L_{\mathbf{W}\mathbf{\Sigma}}(\mathbf{x}))^2 \big] + 2\mathbb{E}_{\mathbf{\Sigma},\mathcal{D}}\Big[(\Delta^Q_{\mathbf{W}\mathbf{\Sigma}}(\mathbf{x}))^2 \Big] \Big)^{1/2} \\
&= \widetilde{O}\Big(B_x^3 \|\mathbf{W}\|_{2,4}^3 m^{-1/4} + d^2 B_x^2 \|\mathbf{W}\|_{2,4}^2 m^{-1/2} \Big),
\end{aligned}$$

where the last step uses Lemma 11. □

Lemma 13 (Closeness of directional gradients)**.** *We have*

$$\begin{aligned}
&\big| \langle \nabla L(\mathbf{W}), \mathbf{W} \rangle - \big\langle \nabla L^Q(\mathbf{W}), \mathbf{W} \big\rangle \big| \\
&\le \widetilde{O}\Big(\Big(dB_x \|\mathbf{W}\|_{2,4} + \sqrt{d} B_x^5 \|\mathbf{W}\|_{2,4}^5 + B_x^3 \|\mathbf{W}\|_{2,4}^3 \Big) m^{-1/4} + d^{2.5} B_x^4 \|\mathbf{W}\|_{2,4}^4 m^{-1/2} \Big).
\end{aligned}$$

Proof. Differentiating L and L^Q and taking the inner product with $\mathbf{W}$, we get

$$\langle \nabla L(\mathbf{W}), \mathbf{W} \rangle = \mathbb{E}_{\mathbf{\Sigma},\mathcal{D}}\left[\ell'(y, f_{\mathbf{W}_0 + \mathbf{W}\mathbf{\Sigma}}(\mathbf{x})) \cdot \frac{1}{\sqrt{m}} \sum_{r\le m} a_r \sigma'((\mathbf{w}_{0,r} + \Sigma_{rr}\mathbf{w}_r)^\top \mathbf{x})(\Sigma_{rr}\mathbf{w}_r^\top \mathbf{x}) \right],$$

and

$$\big\langle \nabla L^Q(\mathbf{W}), \mathbf{W} \big\rangle = \mathbb{E}_{\mathcal{D}}\left[\ell'(y, f^Q_{\mathbf{W}}(\mathbf{x})) \cdot \frac{1}{\sqrt{m}} \sum_{r\le m} a_r \sigma''(\mathbf{w}_{0,r}^\top \mathbf{x}) \cdot (\mathbf{w}_r^\top \mathbf{x})^2 \right].$$

Therefore, by expanding $\sigma'((\mathbf{w}_{0,r}+\Sigma_{rr}\mathbf{w}_r)^\top\mathbf{x})$ and noticing that $\Sigma_{rr}^2 \equiv 1$, we have

$$\begin{aligned}
\left|\left\langle \nabla L(\mathbf{W}) - \nabla L^Q(\mathbf{W}), \mathbf{W}\right\rangle\right| = \Bigg| & \underbrace{\mathbb{E}_{\boldsymbol{\Sigma},\mathcal{D}}\left[\ell'(y, f_{\mathbf{W}_0+\mathbf{W}\boldsymbol{\Sigma}}(\mathbf{x}))\cdot\frac{1}{\sqrt{m}}\sum_{r\le m} a_r\sigma'(\mathbf{w}_{0,r}^\top\mathbf{x})(\Sigma_{rr}\mathbf{w}_r^\top\mathbf{x})\right]}_{\mathrm{I}} \\
& + \underbrace{\mathbb{E}_{\boldsymbol{\Sigma},\mathcal{D}}\left[\Big(\ell'(y, f_{\mathbf{W}_0+\mathbf{W}\boldsymbol{\Sigma}}(\mathbf{x})) - \ell'(y, f^Q_{\mathbf{W}}(\mathbf{x}))\Big)\cdot\frac{1}{\sqrt{m}}\sum_{r\le m} a_r\sigma''(\mathbf{w}_{0,r}^\top\mathbf{x})\cdot(\mathbf{w}_r^\top\mathbf{x})^2\right]}_{\mathrm{II}} \\
& + \underbrace{\begin{aligned}\mathbb{E}_{\boldsymbol{\Sigma},\mathcal{D}}\Bigg[&\ell'(y, f_{\mathbf{W}_0+\mathbf{W}\boldsymbol{\Sigma}}(\mathbf{x})) \\ &\cdot\frac{1}{\sqrt{m}}\sum_{r\le m} a_r\big(\sigma'((\mathbf{w}_{0,r}+\Sigma_{rr}\mathbf{w}_r)^\top\mathbf{x}) - \sigma'(\mathbf{w}_{0,r}^\top\mathbf{x}) - \sigma''(\mathbf{w}_{0,r}^\top\mathbf{x})(\Sigma_{rr}\mathbf{w}_r^\top\mathbf{x})\big)(\Sigma_{rr}\mathbf{w}_r^\top\mathbf{x})\Bigg]\end{aligned}}_{\mathrm{III}}\Bigg|.
\end{aligned}$$

We now bound the three terms separately. Recall that $|\ell'| \le 1$ and $\ell'(y,z)$ is 1-Lipschitz in z. For term I we have by Cauchy-Schwarz that

$$\begin{aligned}
|\mathrm{I}| &\le \left(\mathbb{E}_{\mathcal{D}}\left[\frac{1}{m}\sum_{r\le m} a_r^2\sigma'(\mathbf{w}_{0,r}^\top\mathbf{x})^2(\mathbf{w}_r^\top\mathbf{x})^2\right]\right)^{1/2} \\
&\le \left(C\max_{r\in[m],i\in[n]}(\mathbf{w}_{0,r}^\top\mathbf{x}_i)^4\cdot\frac{1}{m}\sum_{r\le m}\left\|\mathbf{w}_r\right\|_2^2 B_x^2\right)^{1/2} \\
&\le \widetilde{O}(dB_x)\cdot\left(\frac{1}{m}\sum_{r\le m}\left\|\mathbf{w}_r\right\|_2^4\right)^{1/4} = O\Big(dB_x\left\|\mathbf{W}\right\|_{2,4} m^{-1/4}\Big).
\end{aligned}$$

For term II, we have

$$\begin{aligned}
|\mathrm{II}| &\overset{(i)}{\le} \Big(\mathbb{E}_{\boldsymbol{\Sigma},\mathcal{D}}\Big[(f_{\mathbf{W}_0+\mathbf{W}\boldsymbol{\Sigma}}(\mathbf{x}) - f^Q_{\mathbf{W}}(\mathbf{x}))^2\Big]\Big)^{1/2}\cdot\Big(\mathbb{E}_{\mathcal{D}}\Big[(f^Q_{\mathbf{W}}(\mathbf{x}))^2\Big]\Big)^{1/2} \\
&\overset{(ii)}{\le} \widetilde{O}\Big(B_x^3\left\|\mathbf{W}\right\|_{2,4}^3 m^{-1/4} + d^2B_x^2\left\|\mathbf{W}\right\|_{2,4}^2 m^{-1/2}\Big)\cdot\widetilde{O}(\sqrt{d}B_x^2\left\|\mathbf{W}\right\|_{2,4}^2) \\
&= \widetilde{O}\Big(\sqrt{d}B_x^5\left\|\mathbf{W}\right\|_{2,4}^5 m^{-1/4} + d^{2.5}B_x^4\left\|\mathbf{W}\right\|_{2,4}^4 m^{-1/2}\Big).
\end{aligned}$$

where (i) uses Cauchy-Schwarz and (ii) uses the bounds in Lemma 10 and 11. For term III we first note by the smoothness of σ' that

$$\begin{aligned}
&\left|a_r\big(\sigma'((\mathbf{w}_{0,r}+\Sigma_{rr}\mathbf{w}_r)^\top\mathbf{x}) - \sigma'(\mathbf{w}_{0,r}^\top\mathbf{x}) - \sigma''(\mathbf{w}_{0,r}^\top\mathbf{x})(\Sigma_{rr}\mathbf{w}_r^\top\mathbf{x})\big)(\Sigma_{rr}\mathbf{w}_r^\top\mathbf{x})\right| \\
&\le C\left|\Sigma_{rr}\mathbf{w}_r^\top\mathbf{x}\right|^3 \le CB_x^3\left\|\mathbf{w}_r\right\|_2^3.
\end{aligned}$$

Substituting this bound into term III yields

$$\begin{aligned}
|\mathrm{III}| &\le \frac{1}{\sqrt{m}}\sum_{r\le m} CB_x^3\left\|\mathbf{w}_r\right\|_2^3 \le C\sqrt{m}B_x^3\cdot\frac{1}{m}\sum_{r\le m}\left\|\mathbf{w}_r\right\|_2^3 \\
&\le C\sqrt{m}B_x^3\cdot\left(\frac{1}{m}\sum_{r\le m}\left\|\mathbf{w}_r\right\|_2^4\right)^{3/4} = O\Big(B_x^3\left\|\mathbf{W}\right\|_{2,4}^3 m^{-1/4}\Big).
\end{aligned}$$

Putting together the bounds for term I, II, III gives the desired result. □

Lemma 14 (Closeness of Hessians)**.** *Let $\mathbf{\Sigma}'$ denote a diagonal matrix with diagonal entries drawn i.i.d. from* $\mathrm{Unif}\{\pm 1\}$*. We have for all $\mathbf{W}, \mathbf{W}_\star \in \mathbb{R}^{d\times m}$ that*

$$\begin{aligned}&\left|\mathbb{E}_{\mathbf{\Sigma}'}\left[\left(\nabla^2 L(\mathbf{W}) - \nabla^2 L^Q(\mathbf{W})\right)[\mathbf{W}_\star\mathbf{\Sigma}', \mathbf{W}_\star\mathbf{\Sigma}']\right]\right| \\ &\le \widetilde{O}\Bigg(\left(B_x^3\left\|\mathbf{W}\right\|_{2,4} + \sqrt{d}B_x^5\left\|\mathbf{W}\right\|_{2,4}^3\right)\left\|\mathbf{W}_\star\right\|_{2,4}^2 m^{-1/4} \\ &\qquad + \left(d^{2.5}B_x^4\left\|\mathbf{W}\right\|_{2,4}^2\left\|\mathbf{W}_\star\right\|_{2,4}^2 + B_x^2(d^2 + \left\|\mathbf{W}\right\|_{2,\infty}^4 B_x^4)\left\|\mathbf{W}_\star\right\|_{2,4}^2\right)m^{-1/2} \\ &\qquad + dB_x^4\left\|\mathbf{W}\right\|_{2,4}^2\left\|\mathbf{W}_\star\right\|_{2,4}^2 m^{-1}\Bigg).\end{aligned}$$

Proof. Differentiating L and L^Q twice on the direction $\mathbf{W}_\star\mathbf{\Sigma}'$, we get

$$\begin{aligned}&\nabla^2 L(\mathbf{W})[\mathbf{W}_\star\mathbf{\Sigma}', \mathbf{W}_\star\mathbf{\Sigma}'] \\ &= \underbrace{\mathbb{E}_{\mathbf{\Sigma},\mathbf{\Sigma}',\mathcal{D}}\left[\ell''(y, f_{\mathbf{W}_0+\mathbf{W}\mathbf{\Sigma}}(\mathbf{x}))\cdot\left(\frac{1}{\sqrt{m}}\sum_{r\le m} a_r\sigma'((\mathbf{w}_{0,r}+\Sigma_{rr}\mathbf{w}_r)^\top\mathbf{x})(\Sigma_{rr}\Sigma'_{rr}\mathbf{w}_{\star,r}^\top\mathbf{x})\right)^2\right]}_{\mathrm{I}(L)} \\ &\quad + \underbrace{\mathbb{E}_{\mathbf{\Sigma},\mathcal{D}}\left[\ell'(y, f_{\mathbf{W}_0+\mathbf{W}\mathbf{\Sigma}}(\mathbf{x}))\cdot\frac{1}{\sqrt{m}}\sum_{r\le m} a_r\sigma''((\mathbf{w}_{0,r}+\Sigma_{rr}\mathbf{w}_r)^\top\mathbf{x})(\Sigma_{rr}\Sigma'_{rr}\mathbf{w}_{\star,r}^\top\mathbf{x})^2\right]}_{\mathrm{II}(L)},\end{aligned}$$

and

$$\begin{aligned}\nabla^2 L^Q(\mathbf{W})[\mathbf{W}_\star\mathbf{\Sigma}', \mathbf{W}_\star\mathbf{\Sigma}'] &= \underbrace{\mathbb{E}_{\mathbf{\Sigma}',\mathcal{D}}\left[\ell''(y, f^Q_{\mathbf{W}}(\mathbf{x}))\cdot\left(\frac{1}{\sqrt{m}}\sum_{r\le m} a_r\sigma''(\mathbf{w}_{0,r}^\top\mathbf{x})(\mathbf{w}_r^\top\mathbf{x})(\Sigma'_{rr}\mathbf{w}_{\star,r}^\top\mathbf{x})\right)^2\right]}_{\mathrm{I}(L^Q)} \\ &\quad + \underbrace{\mathbb{E}_{\mathbf{\Sigma}',\mathcal{D}}\left[\ell'(y, f^Q_{\mathbf{W}}(\mathbf{x}))\cdot\frac{1}{\sqrt{m}}\sum_{r\le m} a_r\sigma''(\mathbf{w}_{0,r}^\top\mathbf{x})(\Sigma'_{rr}\mathbf{w}_{\star,r}^\top\mathbf{x})^2\right]}_{\mathrm{II}(L^Q)}.\end{aligned}$$

We first bound the terms $\mathrm{I}(L)$ and $\mathrm{I}(L^Q)$. We have

$$\begin{aligned}\mathrm{I}(L) &= 2\mathbb{E}_{\mathbf{\Sigma},\mathbf{\Sigma}',\mathcal{D}}\left[\frac{1}{m}\sum_{r\le m} a_r^2\sigma'((\mathbf{w}_{0,r}+\Sigma_{rr}\mathbf{w}_r)^\top\mathbf{x})^2(\mathbf{w}_{\star,r}^\top\mathbf{x})^2\right] \\ &\le C\cdot\sup_{\left\|\mathbf{x}\right\|_2=B_x}\frac{1}{m}\sum_{r\le m}\left((\mathbf{w}_{0,r}+\Sigma_{rr}\mathbf{w}_r)^\top\mathbf{x}\right)^4(\mathbf{w}_{\star,r}^\top\mathbf{x})^2 \\ &\le CB_x^2\cdot\frac{1}{m}\sum_{r\le m}(\widetilde{O}(d^2) + \left\|\mathbf{w}_r\right\|_2^4 B_x^4)\left\|\mathbf{w}_{\star,r}\right\|_2^2 \\ &\le \widetilde{O}\Big(B_x^2\Big(d^2 + \left\|\mathbf{W}\right\|_{2,\infty}^4 B_x^4\Big)\left\|\mathbf{W}_\star\right\|_{2,4}^2 m^{-1/2}\Big).\end{aligned}$$

Using similar arguments on $\mathrm{I}(L^Q)$ gives the bound

$$\mathrm{I}(L^Q) \le \widetilde{O}\Big(dB_x^4\left\|\mathbf{W}\right\|_{2,4}^2\left\|\mathbf{W}_\star\right\|_{2,4}^2 m^{-1}\Big). \tag{12}$$

We now shift attention to bounding $\mathrm{II}(L) - \mathrm{II}(L^Q)$. First note that

$$\begin{aligned}&\left|\Delta_r^{\sigma''}(\mathbf{x})\right| := \left|a_r(\sigma''((\mathbf{w}_{0,r}+\Sigma_{rr}\mathbf{w}_r)^\top\mathbf{x}) - \sigma''(\mathbf{w}_{0,r}^\top\mathbf{x}))(\mathbf{w}_{\star,r}^\top\mathbf{x})^2\right| \\ &\le C|\Sigma_{rr}\mathbf{w}_r^\top\mathbf{x}|\cdot(\mathbf{w}_{\star,r}^\top\mathbf{x})^2 \le CB_x^3\left\|\mathbf{w}_r\right\|_2\left\|\mathbf{w}_{\star,r}\right\|_2^2.\end{aligned}$$

Then we have, by applying the bounds in Lemma 10 and 11,

$$
\begin{aligned}
&\left|\mathrm{II}(L) - \mathrm{II}(L^Q)\right| \\
&= \left| \mathbb{E}_{\mathbf{\Sigma},\mathcal{D}}\left[\ell'(y, f_{\mathbf{W}_0+\mathbf{W}\mathbf{\Sigma}}(\mathbf{x})) \cdot \frac{1}{\sqrt{m}} \sum_{r\le m} \Delta_r^{\sigma''}(\mathbf{x}) \right] + \mathbb{E}_{\mathbf{\Sigma},\mathcal{D}}\Big[\Big(\ell'(y, f^Q_{\mathbf{W}_0+\mathbf{W}\mathbf{\Sigma}}(\mathbf{x})) - \ell'(y, f^Q_{\mathbf{W}}(\mathbf{x})) \Big) \cdot 2 f^Q_{\mathbf{W}_\star}(\mathbf{x}) \Big] \right| \\
&\le C \cdot \frac{1}{\sqrt{m}} \sum_{r\le m} B_x^3 \left\|\mathbf{w}_r\right\|_2 \left\|\mathbf{w}_{\star,r}\right\|_2^2 + C \sqrt{ \mathbb{E}\Big[\Big(f^L_{\mathbf{W}\mathbf{\Sigma}}(\mathbf{x}) + \Delta^Q_{\mathbf{W}\mathbf{\Sigma}}(\mathbf{x}) \Big)^2 \Big] } \cdot \Big(\mathbb{E}\Big[f^Q_{\mathbf{W}_\star}(\mathbf{x})^2 \Big] \Big)^{1/2} \\
&\le \widetilde{O}\Big(B_x^3 \left\|\mathbf{W}\right\|_{2,4} \left\|\mathbf{W}_\star\right\|_{2,4}^2 m^{-1/4} \Big) + \widetilde{O}\Big(B_x^3 \left\|\mathbf{W}\right\|_{2,4}^3 m^{-1/4} + d^2 B_x^2 \left\|\mathbf{W}\right\|_{2,4}^2 m^{-1/2} \Big) \cdot \widetilde{O}\Big(\sqrt{d} B_x^2 \left\|\mathbf{W}_\star\right\|_{2,4}^2 \Big). \\
&= \widetilde{O}\Big(\Big(B_x^3 \left\|\mathbf{W}\right\|_{2,4} + \sqrt{d} B_x^5 \left\|\mathbf{W}\right\|_{2,4}^3 \Big) \left\|\mathbf{W}_\star\right\|_{2,4}^2 m^{-1/4} + d^{2.5} B_x^4 \left\|\mathbf{W}\right\|_{2,4}^2 \left\|\mathbf{W}_\star\right\|_{2,4}^2 m^{-1/2} \Big).
\end{aligned}
$$

Combining all the bounds gives the desired result. □

B.4 Proof of Theorem 2

We apply Lemma 12, 13, and 14 to connect the neural net loss L to the "clean risk" L^Q. First, by Lemma 12, we have for all the assumed $\mathbf{W}$ that

$$
\left|L(\mathbf{W}) - L^Q(\mathbf{W})\right| \le \widetilde{O}\Big(B_x^3 \left\|\mathbf{W}\right\|_{2,4}^3 m^{-1/4} + d^2 B_x^2 \left\|\mathbf{W}\right\|_{2,4}^2 m^{-1/2} \Big).
$$

Therefore we have $\left|L(\mathbf{W}) - L^Q(\mathbf{W})\right| \le \epsilon/6$ so long as

$$
m \ge \widetilde{O}\big(B_x^{12} B_w^{12} \epsilon^{-4} + d^4 B_x^4 B_w^4 \epsilon^{-2} \big). \tag{13}
$$

Applying Lemma 1, we obtain that

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{\Sigma}'}\left[\nabla^2 L^Q(\mathbf{W})[\mathbf{W}_\star \mathbf{\Sigma}', \mathbf{W}_\star \mathbf{\Sigma}'] \right] - \left\langle \nabla L^Q(\mathbf{W}), \mathbf{W} \right\rangle \\
&\le 2(L^Q(\mathbf{W}) - \mathsf{OPT}) + \epsilon/3 \le 2(L(\mathbf{W}) - \mathsf{OPT}) + 2\epsilon/3
\end{aligned} \tag{14}
$$

provided that the error term in Lemma 1 is bounded by $\epsilon/3$, which happens when

$$
m \ge \widetilde{O}\big(d B_x^4 B_w^2 B_{w,\star}^2 \epsilon^{-1} \big). \tag{15}
$$

Finally, we choose m sufficiently large so that

$$
\left| \mathbb{E}_{\mathbf{\Sigma}'}\left[(\nabla^2 L(\mathbf{W}) - \nabla^2 L^Q(\mathbf{W}))[\mathbf{W}_\star \mathbf{\Sigma}', \mathbf{W}_\star \mathbf{\Sigma}'] \right] \right| \le \epsilon/6
$$

and

$$
\left| \left\langle \nabla L(\mathbf{W}) - \nabla L^Q(\mathbf{W}), \mathbf{W} \right\rangle \right| \le \epsilon/6,
$$

which combined with (14) yields the desired result. By Lemma 13 and 14, it suffices to choose m such that, to satisfy the closeness of directional gradients,

$$
m \ge \widetilde{O}\big((d^4 B_x^4 B_w^4 + d^2 B_x^{20} B_w^{20} + B_x^{12} B_w^{12}) \epsilon^{-4} + d^5 B_x^8 B_w^8 \epsilon^{-2} \big), \tag{16}
$$

and to satisfy the closeness of Hessian quadratic forms,

$$
\begin{aligned}
m \ge O\bigg(&\big[B_x^{12} B_w^4 B_{w,\star}^8 + d^2 B_x^{20} B_w^{12} B_{w,\star}^8 \big] \epsilon^{-4} \\
&+ \big[d^5 B_x^8 B_w^4 B_{w,\star}^4 + d^4 B_x^4 B_{w,\star}^4 + B_x^{12} B_w^8 B_{w,\star}^4 \big] \epsilon^{-2} + d B_x^4 B_w^2 B_{w,\star}^2 \epsilon^{-1} \bigg).
\end{aligned} \tag{17}
$$

Collecting the requirements on m in (13), (15), (16), (17) and merging terms using $\epsilon \le 1$ and $B_{w,\star} \le B_w$, the desired result holds whenever

$$
m \ge O\big(\big[B_x^{12} B_w^{12} + d^4 B_x^4 B_w^4 + d^2 B_x^{20} B_w^{20} \big] \epsilon^{-4} + d^5 B_x^8 B_w^8 \epsilon^{-2} \big).
$$

This completes the proof. □

B.5 Proof of Corollary 3

Proof. For all $\lambda \ge 0$ define

$$A_\lambda := \mathbb{E}_{\mathbf{\Sigma}'}\left[\nabla^2 L_\lambda(\mathbf{W})[\mathbf{W}_\star\mathbf{\Sigma}', \mathbf{W}_\star\mathbf{\Sigma}']\right] - \langle \nabla L_\lambda(\mathbf{W}), \mathbf{W}\rangle + 2(L_\lambda(\mathbf{W}) - \mathsf{OPT}),$$

By Lemma 2, it suffices to show that

$$A_\lambda - A_0 \le C\lambda \|\mathbf{W}_\star\|_{2,4}^8 - \lambda \|\mathbf{W}\|_{2,4}^8$$

for some absolute constant C.

Recall that $L_\lambda(\mathbf{W}) = L(\mathbf{W}) + \lambda \|\mathbf{W}_\star\|_{2,4}^8$. By differentiating $\mathbf{A} \mapsto \|\mathbf{A}\|_{2,4}^8$ we get

$$\langle \nabla (L_\lambda - L)(\mathbf{W}), \mathbf{W}\rangle = 2\|\mathbf{W}\|_{2,4}^4 \cdot \sum_{r\le m} \left\langle 4\lambda \|\mathbf{w}_r\|_2^2 \mathbf{w}_r, \mathbf{w}_r \right\rangle = 8\lambda \|\mathbf{W}\|_{2,4}^8$$

and

$$\begin{aligned}
&\nabla^2 (L_\lambda - L)(\mathbf{W})[\mathbf{W}_\star\mathbf{\Sigma}', \mathbf{W}_\star\mathbf{\Sigma}'] \\
&= 8\lambda \|\mathbf{W}\|_{2,4}^4 \left[\sum_{r\le m} \|\mathbf{w}_r\|_2^2 \|\mathbf{w}_{\star,r}\Sigma'_{rr}\|_2^2 + 2\langle \mathbf{w}_r, \mathbf{w}_{\star,r}\Sigma'_{rr}\rangle^2\right] + 32\lambda \|\mathbf{W}\|_{2,4}^4 \cdot \left(\sum_{r\le m} \|\mathbf{w}_r\|_2^2 \langle \mathbf{w}_r, \mathbf{w}_{\star,r}\Sigma'_{rr}\rangle\right)^2 \\
&\le 56\lambda \|\mathbf{W}\|_{2,4}^4 \sum_{r\le m} \|\mathbf{w}_r\|_2^2 \|\mathbf{w}_{\star,r}\|_2^2 \overset{(i)}{\le} 56\lambda \|\mathbf{W}\|_{2,4}^6 \|\mathbf{W}_\star\|_{2,4}^2 \overset{(ii)}{\le} 14\lambda\alpha \|\mathbf{W}\|_{2,4}^8 + \frac{378\lambda}{\alpha^3}\|\mathbf{W}_\star\|_{2,4}^8,
\end{aligned}$$

where (i) used Cauchy-Schwarz and (ii) used the AM-GM inequality $p^3 q \le \alpha p^4/4 + 27q^4/(4\alpha^3)$ for all p, q and $\alpha > 0$. Substituting the above expressions into $A_\lambda - A_0$ yields

$$\begin{aligned}
&A_\lambda - A_0 \\
&\le 14\lambda\alpha \|\mathbf{W}\|_{2,4}^8 + \frac{378\lambda}{\alpha^3}\|\mathbf{W}_\star\|_{2,4}^8 - 8\lambda \|\mathbf{W}\|_{2,4}^8 + 2\lambda \|\mathbf{W}\|_{2,4}^8 \\
&= (14\lambda\alpha - 6\lambda)\|\mathbf{W}\|_{2,4}^8 + \frac{378\lambda}{\alpha^3}\|\mathbf{W}_\star\|_{2,4}^8.
\end{aligned}$$

Choosing $\alpha = 5/14$ gives the desired result. □

B.6 Proof of Theorem 4

We begin by choosing the regularization strength as

$$\lambda = \lambda_0 B_{w,\star}^{-8},$$

where λ_0 is a constant to be determined. Let ϵ be an accuracy parameter also to be determined.

Localizing second-order stationary points We first argue that any second order stationary point $\mathbf{W}$ has to satisfy $\|\mathbf{W}\|_{2,4} \le B_{w,0}$ for some large but controlled $B_{w,0}$. We first note that for the clean risk L^Q, we have for any $\mathbf{W} \in \mathbb{R}^{d\times m}$ that

$$\begin{aligned}
&\left\langle \nabla L^Q(\mathbf{W}), \mathbf{W}\right\rangle = \mathbb{E}_{\mathcal{D}}\left[\ell'(y, f^Q_{\mathbf{W}}(\mathbf{x})) \cdot 2 f^Q_{\mathbf{W}}(\mathbf{x})\right] \\
&= 2\mathbb{E}_{\mathcal{D}}\left[\ell'(y, f^Q_{\mathbf{W}}(\mathbf{x})) \cdot (f^Q_{\mathbf{W}}(\mathbf{x}) - f^Q_{\mathbf{0}}(\mathbf{x}))\right] \overset{(i)}{\ge} 2(L^Q(\mathbf{W}) - L^Q(\mathbf{0})) \overset{(ii)}{\ge} -2,
\end{aligned}$$

where (i) uses convexity of ℓ and (ii) uses the assumption that $\ell(y, 0) \le 1$ for all $y \in \mathcal{Y}$.

Now, applying the coupling Lemma 13, and combining with the fact that $\left\langle \nabla_{\mathbf{W}}(\lambda \|\mathbf{W}\|_{2,4}^8), \mathbf{W}\right\rangle = 8\lambda \|\mathbf{W}\|_{2,4}^8$, we have simultaneously for all $\mathbf{W}$ that

$$\begin{aligned}
&\langle \nabla L_\lambda(\mathbf{W}), \mathbf{W}\rangle \\
&\ge \left\langle \nabla_{\mathbf{W}}(\lambda \|\mathbf{W}\|_{2,4}^8), \mathbf{W}\right\rangle + \langle \nabla L^Q(\mathbf{W}), \mathbf{W}\rangle - \left|\langle \nabla (L - L^Q)(\mathbf{W}), \mathbf{W}\rangle\right| \\
&\ge 8\lambda \|\mathbf{W}\|_{2,4}^8 - 2 - \widetilde{O}\Big(\Big(dB_x \|\mathbf{W}\|_{2,4} + \sqrt{d}B_x^5 \|\mathbf{W}\|_{2,4}^5 + B_x^3 \|\mathbf{W}\|_{2,4}^3\Big)m^{-1/4} + d^{2.5}B_x^4 \|\mathbf{W}\|_{2,4}^4 m^{-1/2}\Big).
\end{aligned}$$

Therefore we see that any stationary point $\mathbf{W}$ has to satisfy

$$\begin{aligned} &\|\mathbf{W}\|_{2,4} \le B_{w,0} \\ &:= \widetilde{O}\Big(\lambda^{-1/8} + (\lambda^{-1} d B_x m^{-1/4})^{1/7} + (\lambda^{-1}\sqrt{d} B_x^5 m^{-1/4})^{1/3} + (\lambda^{-1} B_x^3)^{1/5} + (\lambda^{-1} d^{2.5} B_x^4 m^{-1/2})^{1/4}\Big). \end{aligned}$$

By Corollary 3, choosing $m \ge \mathrm{poly}(\lambda_0^{-1}, d, B_{w,\star} B_x, \epsilon)$, the coupling error is bounded by ϵ in $\mathsf{B}_{2,4}(B_{w,0})$, i.e. for all $\mathbf{W} \in \mathsf{B}_{2,4}(B_{w,0})$ we have that

$$\begin{aligned} &\mathbb{E}_{\mathbf{\Sigma}'}\left[\nabla^2 L_\lambda(\mathbf{W})[\mathbf{W}_\star \mathbf{\Sigma}', \mathbf{W}_\star \mathbf{\Sigma}']\right] \\ &\le \langle \nabla L_\lambda(\mathbf{W}), \mathbf{W} \rangle - 2(L_\lambda(\mathbf{W}) - \mathsf{OPT}) - \lambda \|\mathbf{W}\|_{2,4}^8 + C\lambda \|\mathbf{W}_\star\|_{2,4}^8 + \epsilon, \end{aligned} \tag{18}$$

where $C = O(1)$ is an absolute constant.

Bounding loss and norm Choosing

$$\lambda_0 = \frac{1}{C} \cdot (2\gamma \mathsf{OPT} + \epsilon),$$

we get that $C\lambda B_{w,\star}^8 = 2\gamma\mathsf{OPT} + \epsilon$, and thus the bound (18) reads

$$\begin{aligned} &\mathbb{E}_{\mathbf{\Sigma}'}\left[\nabla^2 L_\lambda(\mathbf{W})[\mathbf{W}_\star \mathbf{\Sigma}', \mathbf{W}_\star \mathbf{\Sigma}']\right] \\ &\le \langle \nabla L_\lambda(\mathbf{W}), \mathbf{W} \rangle - 2(L_\lambda(\mathbf{W}) - \mathsf{OPT}) - \lambda \|\mathbf{W}\|_{2,4}^8 + 2\gamma\mathsf{OPT} + 2\epsilon. \end{aligned}$$

For the second-order stationary point $\widehat{\mathbf{W}}$, the gradient term vanishes and the Hessian term is non-negative, so we get

$$2(L_\lambda(\widehat{\mathbf{W}}) - \mathsf{OPT}) \le 2(\gamma\mathsf{OPT} + \epsilon) - \lambda \left\|\widehat{\mathbf{W}}\right\|_{2,4}^8 \le 2(\gamma\mathsf{OPT} + \epsilon)$$

and thus

$$L_\lambda(\widehat{\mathbf{W}}) \le (1+\gamma)\mathsf{OPT} + \epsilon.$$

Further, by re-writing (18), we obtain

$$\begin{aligned} \lambda \left\|\widehat{\mathbf{W}}\right\|_{2,4}^8 &\le C\lambda B_{w,\star}^8 + 2(\mathsf{OPT} - L_\lambda(\widehat{\mathbf{W}})) + \epsilon \le C\lambda B_{w,\star}^8 + 2\mathsf{OPT} + \epsilon \\ &\le C\lambda B_{w,\star}^8 \cdot \left(1 + \frac{2\mathsf{OPT} + \epsilon}{2\gamma\mathsf{OPT} + \epsilon}\right) = O(1) \cdot \lambda B_{w,\star}^8, \end{aligned}$$

for any $\gamma = O(1)$. This is the desired result. □

C Proofs for Section 5

C.1 Proof of Lemma 5

As the loss $\ell(y, z)$ is 1-Lipschitz in z for all y, by the Rademacher contraction theorem (Wainwright, 2019, Chapter 5) we have that

$$\begin{aligned} R(\mathcal{F}^Q(B_w)) &\le 2\mathbb{E}_{\boldsymbol{\sigma},\mathbf{x}}\left[\sup_{\|\mathbf{W}\|_{2,4} \le B_w} \frac{1}{n}\sum_{i=1}^n \sigma_i f_{\mathbf{W}}^Q(\mathbf{x}_i)\right] + \mathbb{E}_{\boldsymbol{\sigma},\mathbf{x}}\left[\frac{1}{n}\sum_{i=1}^n \sigma_i \ell(y_i, 0)\right] \\ &\le 2\mathbb{E}_{\boldsymbol{\sigma},\mathbf{x}}\left[\sup_{\|\mathbf{W}\|_{2,4} \le B_w} \frac{1}{\sqrt{m}} \sum_{r \le m} \left\langle \frac{1}{n}\sum_{i=1}^n \sigma_i a_r \sigma''(\mathbf{w}_{0,r}^\top \mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^\top, \mathbf{w}_r\mathbf{w}_r^\top \right\rangle\right] + \frac{1}{\sqrt{n}} \\ &\le 2\mathbb{E}_{\boldsymbol{\sigma},\mathbf{x}}\left[\sup_{\|\mathbf{W}\|_{2,4} \le B_w} \max_{r\in[m]} \left\|\frac{1}{n}\sum_{i=1}^n a_r \sigma''(\mathbf{w}_{0,r}^\top \mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^\top\right\|_{\rm op} \cdot \frac{1}{\sqrt{m}} \sum_{r \le m} \left\|\mathbf{w}_r\mathbf{w}_r^\top\right\|_*\right] + \frac{1}{\sqrt{n}} \\ &\le 2\mathbb{E}_{\boldsymbol{\sigma},\mathbf{x}}\left[\max_{r\in[m]} \left\|\frac{1}{n}\sum_{i=1}^n \sigma''(\mathbf{w}_{0,r}^\top \mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^\top\right\|_{\rm op}\right] \cdot \underbrace{\sup_{\|\mathbf{W}\|_{2,4} \le B_w} \frac{1}{\sqrt{m}} \|\mathbf{w}_r\|_2^2}_{\le B_w^2} + \frac{1}{\sqrt{n}}, \end{aligned}$$

where the last step used the power mean (or Cauchy-Schwarz) inequality on $\{\|\mathbf{w}_r\|_2\}$. □

C.2 Proof of Theorem 6

We first relate the generalization of L to that of L^Q through

$$L_P(\widehat{\mathbf{W}}) - L(\widehat{\mathbf{W}}) \le L_P(\widehat{\mathbf{W}}) - L_P^Q(\widehat{\mathbf{W}}) + L_P^Q(\widehat{\mathbf{W}}) - L^Q(\widehat{\mathbf{W}}) + L^Q(\widehat{\mathbf{W}}) - L(\widehat{\mathbf{W}}).$$

By Lemma 12, we have simultaneously for all $\mathbf{W} \in \mathsf{B}_{2,4}(B_w)$ that

$$\left|L(\mathbf{W}) - L^Q(\mathbf{W})\right| \le \widetilde{O}\Big(B_x^3 B_w^3 m^{-1/4} + d^2 B_x^2 B_w^2 m^{-1/2}\Big). \tag{19}$$

Further, from the proof we see that the argument does not depend on the distribution of $\mathbf{x}$ (it holds uniformly for all $\mathbf{x} \in \mathbb{S}^{d-1}(B_x)$, therefore for the population version we also have the bound

$$\left|L_P(\mathbf{W}) - L_P^Q(\mathbf{W})\right| \le \widetilde{O}\Big(B_x^3 B_w^3 m^{-1/4} + d^2 B_x^2 B_w^2 m^{-1/2}\Big). \tag{20}$$

These bounds hold for all $\mathbf{W} \in \mathsf{B}_{2,4}(B_w)$ so apply to $\widehat{\mathbf{W}}$. Therefore it remains to bound $L_P^Q(\widehat{\mathbf{W}}) - L^Q(\widehat{\mathbf{W}})$, i.e. the generalization of the quadratic model.

Generalization of quadratic model By symmetrization and applying Lemma 5, we have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{W}_0,\mathcal{D}}\Big[L_P^Q(\widehat{\mathbf{W}}) - L^Q(\widehat{\mathbf{W}})\Big] \le \mathbb{E}_{\mathbf{W}_0,\mathcal{D}}\left[\sup_{\|\mathbf{W}\|_{2,4} \le B_w} L_P^Q(\mathbf{W}) - L^Q(\mathbf{W})\right] \\
&\le 2\mathbb{E}_{\mathbf{W}_0,\boldsymbol{\sigma},\mathbf{x}}\left[\sup_{\|\mathbf{W}\|_{2,4} \le B_w} \frac{1}{n}\sum_{i=1}^n \sigma_i \ell(y_i, f_{\mathbf{W}}^Q(\mathbf{x}_i))\right] \\
&\le 4B_w^2 \mathbb{E}_{\mathbf{W}_0,\boldsymbol{\sigma},\mathbf{x}}\left[\max_{r\in[m]}\left\|\frac{1}{n}\sum_{i=1}^n \sigma_i \sigma''(\mathbf{w}_{0,r}^\top \mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^\top\right\|_{\rm op}\right] + \frac{2}{\sqrt{n}}.
\end{aligned} \tag{21}$$

We now focus on bounding the expected max operator norm above. First, we apply the matrix concentration Lemma 8 to deduce that

$$\begin{aligned}
&\mathbb{E}_{\mathbf{W}_0,\boldsymbol{\sigma},\mathbf{x}}\left[\max_{r\in[m]}\left\|\frac{1}{n}\sum_{i=1}^n \sigma_i \sigma''(\mathbf{w}_{0,r}^\top \mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^\top\right\|_{\rm op}\right] \\
&\le 4\sqrt{\log(2dm)} \cdot \mathbb{E}_{\mathbf{W}_0,\mathbf{x}}\left[\sqrt{\max_{r\in[m]}\left\|\frac{1}{n^2}\sum_{i=1}^n \sigma''(\mathbf{w}_{0,r}^\top \mathbf{x}_i)^2 \left\|\mathbf{x}_i\right\|_2^2 \mathbf{x}_i\mathbf{x}_i^\top\right\|_{\rm op}}\right] \\
&\le 4B_x\sqrt{\frac{\log(2dm)}{n}} \cdot \mathbb{E}_{\mathbf{W}_0,\mathbf{x}}\left[\sqrt{\max_{r,i} \sigma''(\mathbf{w}_{0,r}^\top \mathbf{x}_i)^2 \cdot \left\|\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top\right\|_{\rm op}}\right] \\
&\le 4B_x\sqrt{\frac{\log(2dm)}{n}}\left(\mathbb{E}_{\mathbf{W}_0,\mathbf{x}}\left[\max_{r,i}\sigma''(\mathbf{w}_{0,r}^\top \mathbf{x}_i)^2\right] \cdot \mathbb{E}_{\mathbf{x}}\left[\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top\right\|_{\rm op}\right]\right)^{1/2}
\end{aligned}$$

As $|\sigma''(t)| \le Ct^2$ and $\mathbf{w}_{0,r}^\top \mathbf{x}_i \sim \mathsf{N}(0,1)$ for all (r,i), by standard expected max bound on sub-exponential variables we have

$$\mathbb{E}_{\mathbf{W}_0,\mathbf{x}}\left[\max \sigma''(\mathbf{w}_{0,r}^\top \mathbf{x}_i)^2\right] \le O(\log(mn)) = \widetilde{O}(1).$$

Therefore defining

$$M_{x,{\rm op}} := \left(B_x^{-2} \cdot \mathbb{E}_{\mathbf{x}}\left[\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top\right]\right)^{1/2},$$

and substituting the above bound into (21) yields that

$$\mathbb{E}_{\mathbf{W}_0,\mathcal{D}}\Big[L_P^Q(\widehat{\mathbf{W}}) - L^Q(\widehat{\mathbf{W}})\Big] \le \widetilde{O}\left(\frac{B_x^2 B_w^2 M_{x,{\rm op}}}{\sqrt{n}} + \frac{1}{\sqrt{n}}\right).$$

Combining the bound with the coupling error (19) and (20), we arrive at the desired result.

For $M_{x,{\rm op}}$ we have two versions of bounds:

(a) We always have $\left\|\sum_{i\le n}\mathbf{x}_i\mathbf{x}_i^\top/n\right\|_{\rm op}\le B_x^2$, and thus $M_{x,{\rm op}}\le 1$.

(b) If, in addition, $\mathbf{x}$ is uniformly distributed on the sphere $\mathbb{S}^{d-1}(B_x)$ or the hypercube $\left\{\pm B_x/\sqrt{d}\right\}^d$, then we have by standard covariance concentration (Vershynin, 2018, Theorem 4.7.1) that $\mathbb{E}_{\mathbf{x}}\left[\left\|\sum_{i\le n}\mathbf{x}_i\mathbf{x}_i^\top/n\right\|_{\rm op}\right]\le B_x^2/d\cdot O(1+\sqrt{d/n}+d/n)=O(B_x^2/d)$ when $n\ge d$. More generally, if for all $\mathbf{v}\in\mathbb{S}^{d-1}(1)$ we have

$$\left\|\mathbf{v}^\top\mathbf{x}\right\|_{\psi_2}\le K\sqrt{\mathbf{v}^\top{\rm Cov}(\mathbf{x})\mathbf{v}},$$

and that $\kappa({\rm Cov}(\mathbf{x}))\le\kappa$, then we have $\|{\rm Cov}(\mathbf{x})\|_{\rm op}\le\kappa B_x^2/d$. Applying (Vershynin, 2018, Theorem 4.7.1), we get $M_{x,{\rm op}}\le\kappa/\sqrt{d}$ whenever $n\ge O(K^4d)$.

□

C.3 Expressive power of infinitely wide quadratic models

Lemma 15 (Expressivity of f^Q with infinitely many neurons)**.** *Suppose $f_\star(\mathbf{x})=\alpha(\boldsymbol{\beta}^\top\mathbf{x})^p$ for some $\alpha\in\mathbb{R}$, $\boldsymbol{\beta}\in\mathbb{R}^d$, and $p\ge 2$ and such that $p-2\in\{1\}\cup\{2\ell\}_{\ell\ge 0}$. Suppose further that we use $\sigma(t)=\frac{1}{6}{\rm relu}^3(t)$ (so that $\sigma''(t)={\rm relu}(t)$), then there exists choices of $(\mathbf{w}_+,\mathbf{w}_-)$ that depends on $\mathbf{w}_0$ such that*

$$\mathbb{E}_{\mathbf{w}_0}\left[\sigma''(\mathbf{w}_0^\top\mathbf{x})\left((\mathbf{w}_+^\top\mathbf{x})^2-(\mathbf{w}_-^\top\mathbf{x})^2\right)\right]=f_\star(\mathbf{x})$$

and further satisfies the norm bound

$$\mathbb{E}_{\mathbf{w}_0}\left[\|\mathbf{w}_+\|_2^4+\|\mathbf{w}_-\|_2^4\right]\le 2\pi((p-2)\vee 1)^3\alpha^2B_x^{2(p-2)}\|\boldsymbol{\beta}\|_2^{2p}.$$

Proof. Our proof builds on reducing the problem from representing $(\boldsymbol{\beta}^\top\mathbf{x})^p$ via quadratic networks to representing $(\boldsymbol{\beta}^\top\mathbf{x})^{p-2}$ through a random feature model. More precisely, we consider choosing

$$(\mathbf{w}_+,\mathbf{w}_-)=\left(\sqrt{[a]_+}\cdot\boldsymbol{\beta},\sqrt{[a]_-}\cdot\boldsymbol{\beta}\right),\tag{22}$$

where a is a real-valued random scalar that can depend on $\mathbf{w}_0$, and $\boldsymbol{\beta}$ is the fixed coefficient vector in $f_\star$. With this choice, the quadratic network reduces to

$$\begin{aligned}&\mathbb{E}_{\mathbf{w}_0}\left[\sigma''(\mathbf{w}_0^\top\mathbf{x})\left((\mathbf{w}_+^\top\mathbf{x})^2-(\mathbf{w}_-^\top\mathbf{x})^2\right)\right]\\&=\mathbb{E}_{\mathbf{w}_0}\left[\sigma''(\mathbf{w}_0^\top\mathbf{x})\left(a_+(\boldsymbol{\beta}^\top\mathbf{x})^2-a_-(\boldsymbol{\beta}^\top\mathbf{x})^2\right)\right]=(\beta^\top\mathbf{x})^2\mathbb{E}_{\mathbf{w}_0}\left[\sigma''(\mathbf{w}_0^\top\mathbf{x})a\right].\end{aligned}$$

Therefore, to let the above express $f_\star(\mathbf{x})=\alpha(\boldsymbol{\beta}^\top\mathbf{x})^p$, it suffices to choose a such that

$$\mathbb{E}[\sigma''(\mathbf{w}_0^\top\mathbf{x})a]\equiv\alpha(\boldsymbol{\beta}^\top\mathbf{x})^{p-2}\tag{23}$$

for all $\mathbf{x}$. By Lemma 9, there exists $a=a(\mathbf{w}_0)$ satisfying (23) and such that

$$\mathbb{E}_{\mathbf{w}_0}[a^2]\le 2\pi((p-2)\vee 1)^3\alpha^2B_x^{2(p-2)}\|\boldsymbol{\beta}\|_2^{2(p-2)}.$$

Using this a in (22), the quadratic network induced by $(\mathbf{w}_+,\mathbf{w}_-)$ has the desired expressivity, and further satisfies the expected 4th power norm bound

$$\begin{aligned}&\mathbb{E}_{\mathbf{w}_0}[\|\mathbf{w}_+\|_2^4+\|\mathbf{w}_-\|_2^4]\\&=\mathbb{E}_{\mathbf{w}_0}[[a]_+^2+[a]_-^2]\cdot\|\boldsymbol{\beta}\|_2^4=\mathbb{E}_{\mathbf{w}_0}[a^2]\|\boldsymbol{\beta}\|_2^4\le 2\pi((p-2)\vee 1)^3\alpha^2B_x^{2(p-2)}\|\boldsymbol{\beta}\|_2^{2p}.\end{aligned}$$

This is the desired result.

□

C.4 Proof of Theorem 7

We build on the infinite-neuron construction in Lemma 15. Given the symmetric initialization $\{\mathbf{w}_{0,r}\}_{r=1}^m$, for all $r \in [m/2]$, we consider $\mathbf{W}_\star \in \mathbb{R}^{d\times m}$ defined through

$$(\mathbf{w}_{\star,r}, \mathbf{w}_{\star,r+m/2}) = \Big(2m^{-1/4}\mathbf{w}_+(\mathbf{w}_{0,r}), 2m^{-1/4}\mathbf{w}_-(\mathbf{w}_{0,r})\Big),$$

where we recall $(\mathbf{w}_+(\mathbf{w}_0), \mathbf{w}_-(\mathbf{w}_0)) = (\sqrt{a_+(\mathbf{w}_0)}\boldsymbol{\beta}, \sqrt{a_-(\mathbf{w}_0)}\boldsymbol{\beta})$. We then have

$$\begin{aligned} f^Q_{\mathbf{W}_\star}(\mathbf{x}) &= \frac{1}{2\sqrt{m}} \sum_{r\le m/2} \sigma''(\mathbf{w}_{0,r}^\top \mathbf{x})\Big[(\mathbf{w}_{\star,r}^\top\mathbf{x})^2 - (\mathbf{w}_{\star,r+m/2}^\top\mathbf{x})^2\Big] \\ &= \frac{2}{m}\sum_{r\le m/2}\sigma''(\mathbf{w}_{0,r}^\top\mathbf{x})\big[(\mathbf{w}_+(\mathbf{w}_{0,r})^\top\mathbf{x})^2 - (\mathbf{w}_-(\mathbf{w}_{0,r})^\top\mathbf{x})^2\big] \\ &= \left[\frac{1}{m/2}\sum_{r\le m/2}\sigma''(\mathbf{w}_{0,r}^\top\mathbf{x})a(\mathbf{w}_{0,r})\right]\cdot(\boldsymbol{\beta}^\top\mathbf{x})^2. \end{aligned}$$

Bound on $\|\mathbf{W}_\star\|_{2,4}$ As $f_\star(\mathbf{x}) = \alpha(\boldsymbol{\beta}^\top\mathbf{x})^p$, Lemma 15 guarantees that the coefficient $a(\mathbf{w}_0)$ involved above satisfies that

$$R_a^2 := \mathbb{E}_{\mathbf{w}_0}[a(\mathbf{w}_0)^2] \le 2\pi((p-2)\vee 1)^3\alpha^2 B_x^{2(p-2)} \left\|\boldsymbol{\beta}\right\|_2^{2(p-2)}.$$

By Markov inequality, we have with probability at least $1-\delta/2$ that

$$\frac{1}{m/2}\sum_{r\le m} a(\mathbf{w}_{0,r})^2 \le 4\pi((p-2)\vee 1)^3\alpha^2 B_x^{2(p-2)}\left\|\boldsymbol{\beta}\right\|_2^{2(p-2)}\delta^{-1},$$

which yields the bound

$$\begin{aligned} &\left\|\mathbf{W}\right\|_{2,4}^4 = \sum_{r\le m}\left\|\mathbf{w}_{\star,r}\right\|_2^4 \\ &\le \left\|\boldsymbol{\beta}\right\|_2^4\cdot\sum_{r\le m/2}16m^{-1}a(\mathbf{w}_{0,r})^2 = 8\left\|\boldsymbol{\beta}\right\|_2^4\cdot\frac{1}{m/2}\sum_{r\le m/2}a(\mathbf{w}_{0,r})^2 \\ &\le 32\pi[(p-2)^3\vee 1]\alpha^2 B_x^{2(p-2)}\left\|\boldsymbol{\beta}\right\|_2^{2p}\delta^{-1}. \end{aligned}$$

Concentration of function Let $f_m(\mathbf{x}) = \frac{1}{m}\sum_{r\le m/2}\sigma''(\mathbf{w}_{0,r}^\top\mathbf{x})a(\mathbf{w}_{0,r})$. We now show the concentration of f_m to $f_{\star,p-2}(\mathbf{x}) := \alpha(\beta^\top\mathbf{x})^{p-2}$ over the dataset $\{\mathbf{x}_1,\dots,\mathbf{x}_n\}$. We perform a truncation argument: let R be a large radius (to be chosen) satisfying

$$\mathbb{P}_{\mathbf{W}_0}\left(\sup_{r\in[m]}\left\|\mathbf{w}_{0,r}\right\|_2 \ge RB_x^{-1}\right) \ge 1-\delta/2. \tag{24}$$

On this event we have

$$f_m(\mathbf{x}) = \frac{1}{m}\sum_{r\le m}\sigma''(\mathbf{w}_{0,r}^\top\mathbf{x})a(\mathbf{w}_{0,r})\mathbf{1}\left\{\left\|\mathbf{w}_{0,r}\right\|_2 \le RB_x^{-1}\right\} := f_m^R(\mathbf{x}).$$

Letting $f_{\star,p-2}^R(\mathbf{x}) := \mathbb{E}_{\mathbf{w}_0}[\sigma''(\mathbf{w}_0^\top\mathbf{x})a(\mathbf{w}_0)\mathbf{1}\left\{\left\|\mathbf{w}_0\right\|_2 \le RB_x^{-1}\right\}]$, we have

$$\mathbb{E}_{\mathbf{W}_0}\left[\left(f_m(\mathbf{x}) - f_{\star,p-2}^R(\mathbf{x})\right)^2\right] = \frac{1}{m}\mathbb{E}_{\mathbf{w}_0}\left[\sigma''(\mathbf{w}_0^\top\mathbf{x})a^2(\mathbf{w}_0)\mathbf{1}\left\{\left\|\mathbf{w}_0\right\|_2 \le R\right\}\right] \le C\frac{R^2R_a^2}{m}.$$

Applying Chebyshev inequality and a union bound, we get

$$\mathbb{P}\Big(\max_i\left|f_m(\mathbf{x}_i) - f_{\star,p-2}(\mathbf{x}_i)\right| \ge t\Big) \le C\frac{nR^2R_a^2}{mt^2}.$$

For any $\epsilon > 0$, by substituting in $t = \epsilon B_x^{-2} \left\|\boldsymbol{\beta}\right\|_2^{-2}/2$, we see that

$$m \ge O\Big(nR^2 R_a^2 B_x^4 \left\|\boldsymbol{\beta}\right\|_2^4 \epsilon^{-2}\Big) = O\Big(nR^2 (p-2)^3 \alpha^2 B_x^{2p} \left\|\boldsymbol{\beta}\right\|_2^{2p} \epsilon^{-2}\Big) \tag{25}$$

ensures that

$$\max_{i\in[n]} \left|f_m(\mathbf{x}_i) - f_{\star,p-2}^R(\mathbf{x}_i)\right| \le \epsilon B_x^{-2} \left\|\boldsymbol{\beta}\right\|^{-2}/2. \tag{26}$$

Next, for any $\mathbf{x}$ we have the bound

$$\begin{aligned}
&\left|f_{\star,p-2}^R(\mathbf{x}) - f_{\star,p-2}(\mathbf{x})\right| = \left|\mathbb{E}_{\mathbf{w}_0}[\sigma''(\mathbf{w}_0^\top \mathbf{x}) a(\mathbf{w}_0) \mathbf{1}\left\{\left\|\mathbf{w}_0\right\|_2 > R\right\}]\right| \\
&\le \mathbb{E}[a(\mathbf{w}_0)^2]^{1/2} \cdot \mathbb{E}[\sigma''(\mathbf{w}_0^\top \mathbf{x})^4]^{1/4} \cdot \mathbb{P}(\left\|\mathbf{w}_0\right\|_2 > R)^{1/4} \\
&\le R_a \cdot C/\sqrt{d} \cdot \mathbb{P}(\left\|\mathbf{w}_0\right\|_2 > R)^{1/4}.
\end{aligned}$$

Choosing R such that

$$\mathbb{P}(\left\|\mathbf{w}_0\right\|_2 > R) \le c\frac{\sqrt{d}\epsilon^4}{R_a B_x^8 \left\|\boldsymbol{\beta}\right\|_2^8} \tag{27}$$

ensures that

$$\max_i \left|f_{\star,p-2}^R(\mathbf{x}_i) - f_{\star,p-2}(\mathbf{x}_i)\right| \le \frac{\epsilon B_x^{-2} \left\|\boldsymbol{\beta}\right\|_2^{-2}}{2}. \tag{28}$$

Combining (26) and (28), we see that with probability at least $1-\delta$,

$$\begin{aligned}
&\max_{i\in[n]} \left|f_{\mathbf{W}_\star}^Q(\mathbf{x}_i) - f_\star(\mathbf{x}_i)\right| = \max_{i\in[n]} \left|f_m(\mathbf{x}_i) - f_{\star,p-2}(\mathbf{x}_i)\right| \cdot (\boldsymbol{\beta}^\top \mathbf{x}_i)^2 \\
&\le 2 \cdot \frac{\epsilon B_x^{-2} \left\|\boldsymbol{\beta}\right\|_2^{-2}}{2} \cdot B_x^2 \left\|\boldsymbol{\beta}\right\|_2^2 = \epsilon
\end{aligned}$$

and thus

$$\left|L^Q(\mathbf{W}_\star) - L(f_\star)\right| \le \epsilon. \tag{29}$$

To satisfy the requirements for m and R in (27) and (25), we first set $R = \widetilde{O}(\sqrt{d})$ (with sufficiently large log factor) to satisfy (27) by standard Gaussian norm concentration (cf. Appendix A.3), and by (25) it suffices to set m as

$$m \ge \widetilde{O}\big(nd(p-2)^3 \alpha^2 (B_x \left\|\boldsymbol{\beta}\right\|_2)^{2p} \epsilon^{-2}\big).$$

for (29) to hold. □