Limitations

Our main theorem relies on a few assumptions, which are detailed in Section 2. One assumption is that σ satisfies Assumptions 2 and 3, which require σ to be sufficiently differentiable and have nonzero Hermite/Gegenbauer coefficients. While this allows for common activations such as $\mathrm{sigmoid}$ or $\tanh$ (with a generic bias term), it does not include activations such as ReLU. We believe that either considering a smoothed version of ReLU or including a randomized bias term can help us work around this non-differentiability issue. Additionally, we believe the boundedness assumption is not essential and can be overcome with an appropriate truncation argument; however, making the boundedness assumption is useful for simplifying the proof.

Another assumption is that the covariates are drawn from the uniform distribution on the unit sphere $\mathcal{S}^{d-1}(\sqrt{d})$. This assumption is necessary in order to invoke the statistical characterization of the NTK developed in [25, 39], and we note that either the uniform-on-sphere or Gaussian data assumption has been made in a number of prior works [25, 23, 39, 26]. In practice, data can be normalized to be isotropic. Furthermore, neural networks have been observed to generalize better than kernel methods in a wide variety of settings, and thus we believe our results are indicative of a more general phenomenon.

A Spherical Harmonics: Technical Background

Below we present relevant results on spherical harmonics, Gegenbauer polynomials, and Hermite polynomials. These results are from [25, 39], and a more in depth discussion of the technical background can be found in those references.

For $\ell \in \mathbb{Z}^{\geq 0}$, define $B(d,\ell)$ as

$$B(d,\ell) := \frac{2\ell+d-2}{\ell}\binom{\ell+d-3}{\ell-1} = (1+o_d(1))\frac{d^\ell}{\ell!}, \tag{33}$$

and define

$$n_k = \sum_{\ell=0}^{k} B(d,\ell) = \Theta_d(d^k). \tag{34}$$

We let $L^2(\mathcal{S}^{d-1}(\sqrt{d}),\mu)$ denote the space of square-integrable functions over the sphere of radius $\sqrt{d}$, with respect to the uniform probability measure μ. We use the shorthand $\langle\cdot,\cdot\rangle_{L^2} = \langle\cdot,\cdot\rangle_{L^2(\mathcal{S}^{d-1}(\sqrt{d}),\mu)}$, and likewise for $\|f\|_{L^2}$.

The *normalized spherical harmonics* $\{Y_{k,i}^{(d)}\}_{0\leq i\leq B(d,k),k\geq 0}$ are a sequence of polynomials such that $Y_{k,i}^{(d)}$ is degree k, and the $Y_{k,i}^{(d)}$ form an orthonormal basis of $L^2(\mathcal{S}^{d-1}(\sqrt{d}),\mu)$, i.e:

$$\langle Y_{k,i}^{(d)}, Y_{m,j}^{(d)}\rangle_{L^2} = \delta_{km}\delta_{ij}. \tag{35}$$

For $\mathbf{x}\sim\mathcal{S}^{(d-1)}(\sqrt{d})$, let $\tilde{\tau}_{d-1}^1$ be the probability measure of $\sqrt{d}\langle\mathbf{x},\mathbf{e}\rangle$ and τ_{d-1}^1 be the measure of $\langle\mathbf{x},\mathbf{e}\rangle$, where $\mathbf{e}$ is an arbitrary unit vector.

The *Gegenbauer polynomials* $\{Q^{(d)}\}_{k\geq 0}$ are a basis of $L^2([-d,d],\tilde{\tau}_{d-1}^1)$ such that $Q_k^{(d)}$ is a degree k polynomial with $Q_k^{(d)}(d)=1$ and

$$\langle Q_k^{(d)}, Q_j^{(d)}\rangle_{L^2([-d,d],\tilde{\tau}_{d-1}^1)} = \frac{1}{B(d,k)}\delta_{jk}. \tag{36}$$

The following identity relates spherical harmonics and Gegenbauer polynomials:

$$Q_k^{(d)}(\langle\mathbf{x},\mathbf{y}\rangle) = \frac{1}{B(d,k)}\sum_{i=1}^{B(d,k)} Y_{ki}^{(d)}(\mathbf{x})Y_{ki}^{(d)}(\mathbf{y}). \tag{37}$$

We also use the following fact about Gegenbauer polynomials:

$$\langle Q_j^{(d)}(\langle\mathbf{x},\cdot\rangle), Q_k^{(d)}(\langle\mathbf{y},\cdot\rangle)\rangle_{L^2} = \frac{1}{B(d,k)}\delta_{jk}Q_k^{(d)}(\langle\mathbf{x},\mathbf{y}\rangle). \tag{38}$$

Furthermore, $f\in L^2([-\sqrt{d},\sqrt{d}],\tau_{d-1}^1)$ can be decomposed via Gegenbauer polynomials as

$$f(x) = \sum_{k=0}^{\infty}\lambda_{d,k}(\sigma')B(d,k)Q_k^{(d)}(\sqrt{d}x) \tag{39}$$

$$\lambda_{d,k}(f) := \langle f, Q_k^{(d)}(\sqrt{d}\cdot)\rangle_{L^2([-\sqrt{d},\sqrt{d}],\tau_{d-1}^1)}. \tag{40}$$

Thus $\|f\|_{L^2([-\sqrt{d},\sqrt{d}],\tau_{d-1}^1)} = \sum_{k\geq 0} B(d,k)\lambda_{d,k}^2(f)$

Let γ be the measure of a standard Gaussian on $\mathbb{R}$. The *normalized Hermite polynomials* $\{h_k\}_{h\geq 0}$ are an orthonormal basis of $L^2(\mathbb{R},\gamma)$ such that h_k is degree k. For $f\in L^2(\mathbb{R},\gamma)$, let $\mu_k(f) := \langle f, h_k\rangle_{L^2(\mathbb{R},\gamma)}$ be the kth hermite coefficient. One observes that τ_{d-1}^1 converges weakly to γ. As a result we have the following connection between Hermite and Gegenbauer coefficients:

$$\mu_k(f) = \lim_{d\to\infty} B(d,k)^{1/2}\lambda_{d,k}(f). \tag{41}$$

A.1 Assumptions

We additionally require σ', σ'' to satisfy the following assumptions, which are that the Hermite/Gegenbauer coefficients are nonzero and well behaved:

Assumption 3. *We assume σ, σ' satisfy the following:*

(a) Let σ' satisfy $\mu_\ell(\sigma') \neq 0$ for $\ell \leq 4k$ and $\sum_{\ell > 4k} \mu_\ell^2(\sigma') > 0$. As a result, we can let $\mu_\ell^2(\sigma') = \Theta_d(1)$ for $\ell < 4k$.

(b) Let σ'' satisfy

$$d^{k-1} \cdot \min_{\ell \leq k-1} \lambda_{d,\ell}^2(\sigma'') = \Omega_d(1), \tag{42}$$

where $\lambda_{d,\ell}(\sigma'') = \langle \sigma'', Q_\ell^{(d)}(\sqrt{d}\cdot) \rangle_{L^2([-\sqrt{d},\sqrt{d}], \tau_{d-1}^1)}$.

A.2 Computing $\mathbf{\Sigma}$

Below we discuss how to express $\mathbf{\Sigma}$ in terms of the Gegenbauer coefficients of σ', $\lambda_{d,\ell}(\sigma')$. First, observe that that $\mathbf{\Sigma}$ is a matrix of $d \times d$ blocks, where the i,jth block is equal to

$$\frac{a_i a_j}{m} u(\mathbf{w}_{0,i}, \mathbf{w}_{0,j}),$$

where $\mathbf{u} : \mathcal{S}^{d-1}(\sqrt{d}) \times \mathcal{S}^{d-1}(\sqrt{d}) \to \mathbb{R}^{d \times d}$ is the function

$$\mathbf{u}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \mathbb{E}_\mu \left[\sigma'(\boldsymbol{\theta}_1^T \mathbf{x}) \sigma'(\boldsymbol{\theta}_2^T \mathbf{x}) \mathbf{x}\mathbf{x}^T \right].$$

[25, Lemma 7] shows that there exist scalar valued functions u_1, u_2, u_3 such that

$$\mathbf{u}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = u_1(\boldsymbol{\theta}_1^T \boldsymbol{\theta}_2) \mathbf{I}_d + u_2(\boldsymbol{\theta}_1^T \boldsymbol{\theta}_2)[\boldsymbol{\theta}_1 \boldsymbol{\theta}_2^T + \boldsymbol{\theta}_2 \boldsymbol{\theta}_2^T] + u_3(\boldsymbol{\theta}_1^T \boldsymbol{\theta}_2)[\boldsymbol{\theta}_1 \boldsymbol{\theta}_1^T + \boldsymbol{\theta}_2 \boldsymbol{\theta}_2^T],$$

where u_1, u_2, u_3 can be computed in terms of the quantities

$$\text{Tr}(\mathbf{u}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)), \quad \boldsymbol{\theta}_1^T \mathbf{u}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \boldsymbol{\theta}_2, \quad \boldsymbol{\theta}_1^T \mathbf{u}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \boldsymbol{\theta}_1.$$

It thus suffices to compute these quantities. We assume that we can compute arbitrarily many Gegenbauer coefficients of σ'.

Note that

$$\begin{aligned}
\text{Tr}(\mathbf{u}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)) &= d \cdot \mathbb{E}_\mu[\sigma'(\boldsymbol{\theta}_1^T \mathbf{x}) \sigma'(\boldsymbol{\theta}_2^T \mathbf{x})] \\
\boldsymbol{\theta}_1^T \mathbf{u}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \boldsymbol{\theta}_2 &= \mathbb{E}_\mu[\sigma'(\boldsymbol{\theta}_1^T \mathbf{x}) \boldsymbol{\theta}_1^T \mathbf{x} \cdot \sigma'(\boldsymbol{\theta}_2^T \mathbf{x}) \boldsymbol{\theta}_2^T \mathbf{x}] \\
\boldsymbol{\theta}_1^T \mathbf{u}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \boldsymbol{\theta}_1 &= \mathbb{E}_\mu[\sigma'(\boldsymbol{\theta}_1^T \mathbf{x}) (\boldsymbol{\theta}_1^T \mathbf{x})^2 \cdot \sigma'(\boldsymbol{\theta}_2^T \mathbf{x})].
\end{aligned}$$

All these expressions are of the form $\mathbb{E}_\mu[f(\boldsymbol{\theta}_1^T \mathbf{x}) g(\boldsymbol{\theta}_2^T \mathbf{x})]$. For arbitrary f, g, let their decompositions into Gegenbauer polynomials be:

$$f(z) = \sum_{k \geq 0} \lambda_{d,k}(f) B(d,k) Q_k^{(d)}(\sqrt{d}z), \qquad g(z) = \sum_{k \geq 0} \lambda_{d,k}(g) B(d,k) Q_k^{(d)}(\sqrt{d}z).$$

Then, by Equation 38,

$$\begin{aligned}
\mathbb{E}_\mu[f(\boldsymbol{\theta}_1^T \mathbf{x}) g(\boldsymbol{\theta}_2^T \mathbf{x})] &= \sum_{k,\ell \geq 0} \lambda_{d,k}(f) \lambda_{d,\ell}(g) B(d,k) B(d,\ell) \mathbb{E}_\mu \left[Q_k^{(d)}(\sqrt{d} \boldsymbol{\theta}_1^T \mathbf{x}) Q_\ell^{(d)}(\sqrt{d} \boldsymbol{\theta}_2^T \mathbf{x}) \right] \\
&= \sum_{k \geq 0} \lambda_{d,k}(f) \lambda_{d,k}(g) B(d,k) Q_k^{(d)}(d \boldsymbol{\theta}_1^T \boldsymbol{\theta}_2),
\end{aligned}$$

which can be computed to desired precision by truncating this infinite sum accordingly. This only requires knowledge of the Gegenbauer coefficients of f, g. Given the Gegenbauer coefficients of a function $\psi(z)$, [25, Lemma 6] gives a formula for the Gegenbauer coefficients of $z\psi(z)$. We can therefore write the Gegenbauer coefficients of $z\sigma'(z), z^2\sigma'(z)$ in terms of those of σ', and thus we can approximate this sum to the desired precision. This procedure allows us to express $\mathbf{\Sigma}$ in terms of the Gegenbauer coefficients of σ'.

B Expressivity Proofs

B.1 Quad-NTK Proofs

B.1.1 Preliminaries

Lemma 4 (Expressing polynomials with random features)**.** *Let $p \geq 0$, and let σ satisfy the following two assumptions:*

(a) $\sigma \in L^2([-\sqrt{d}, \sqrt{d}], \tau_{d-1}^1)$

(b) $d^p \cdot \min_{k \leq p} \lambda_{d,k}(\sigma) = \Omega_d(1)$*, where* $\lambda_{d,k}(\sigma)^2 = \langle \sigma, Q_k^{(d)}(\sqrt{d}\cdot) \rangle_{L^2([-\sqrt{d},\sqrt{d}], \tau_{d-1}^1)}$.

For $|\alpha| \leq 1$, $\|\beta\| = 1$, there exists a function $a \in L^2(\mathcal{S}^{d-1}(1))$ such that

$$\mathbb{E}_{\mathbf{w}_0 \sim \mathcal{S}^{d-1}(1)} \left[\sigma(\mathbf{w}_0^T \mathbf{x}) a(\mathbf{w}_0) \right] = \alpha(\beta^T \mathbf{x})^p,$$

and a satisfies the norm bound

$$\|a\|^2_{L^2(\mathcal{S}^{d-1}(1))} \lesssim d^p$$

Proof. We can decompose σ into a sum over Gegenbauer polynomials

$$\sigma(x) = \sum_{k=0}^{\infty} \lambda_{d,k}(\sigma) B(d,k) Q_k^{(d)}(\sqrt{d}x),$$

By Equation 37,

$$\sigma(\mathbf{w}_0^T \mathbf{x}) = \sum_{k \geq 0} \sum_{i=1}^{B(d,k)} \lambda_{d,k}(\sigma) Y_{k,i}^{(d)}(\mathbf{x}) Y_{k,i}^{(d)}(\sqrt{d}\mathbf{w}_0).$$

Let a be decomposed into spherical harmonics as

$$a(\mathbf{w}_0) = \sum_{k \geq 0} \sum_{i=1}^{B(d,k)} c_{k,i} Y_{k,i}^{(d)}(\sqrt{d}\mathbf{w}_0),$$

for some coefficients $c_{k,i}$ with $\sum_{k \geq 0} \sum_{i=1}^{B(d,k)} c_{k,i}^2 < \infty$. Since the spherical harmonics form an orthonormal basis of $\mathcal{S}^{d-1}(\sqrt{d})$, we have

$$\mathbb{E}_{\mathbf{w}_0 \sim \mathcal{S}^{d-1}(1)} \left[\sigma(\mathbf{w}_0^T \mathbf{x}) a(\mathbf{w}_0) \right] = \sum_{k \geq 0} \sum_{i=1}^{B(d,k)} \lambda_{d,k}(\sigma) c_{k,i} Y_{k,i}^{(d)}(\mathbf{x}).$$

Next, for an arbitrary function $f \in L^2([-\sqrt{d}, \sqrt{d}], \tau_{d-1}^1)$, we can decompose

$$\begin{aligned} f(\beta^T \mathbf{x}) &= \sum_{k \geq 0} \lambda_{d,k}(f) B(d,k) Q_k(\sqrt{d}\beta^T \mathbf{x}) \\ &= \sum_{k \geq 0} \sum_{i=1}^{B(d,k)} \lambda_{d,k}(f) Y_{k,i}^{(d)}(\mathbf{x}) Y_{k,i}^{(d)}(\sqrt{d}\beta). \end{aligned}$$

For $f(t) = \alpha t^p$, $\lambda_{d,k}(f) = 0$ for $k > p$, and thus

$$\alpha(\beta^T \mathbf{x})^p = \sum_{k=0}^{p} \sum_{i=1}^{B(d,k)} \lambda_{d,k}(f) Y_{k,i}^{(d)}(\mathbf{x}) Y_{k,i}^{(d)}(\sqrt{d}\beta).$$

Define the sequence of coefficients $\{c_{k,i}\}_{0 \leq k \leq p, 1 \leq i \leq B(d,k)}$ by

$$c_{k,i} = Y_{k,i}^{(d)}(\sqrt{d}\beta) \lambda_{d,k}(f) \lambda_{d,k}^{-1}(\sigma),$$

which are well defined for sufficiently large d by assumption (b) of the lemma. Since there are only finitely many nonzero $c_{k,i}$, the function $a(\mathbf{w}_0)$ is in $L_2(\mathcal{S}^{d-1}(1))$. Also,

$$\begin{aligned}\mathbb{E}_{\mathbf{w}_0\sim\mathcal{S}^{d-1}(1)}\left[\sigma(\mathbf{w}_0^T\mathbf{x})a(\mathbf{w}_0)\right] &= \sum_{k\geq 0}\sum_{i=1}^{B(d,k)}\lambda_{d,k}(\sigma)c_{k,i}Y_{k,i}^{(d)}(\mathbf{x})\\ &= \sum_{k=0}^{p}\sum_{i=1}^{B(d,k)}\lambda_{d,k}(f)Y_{k,i}^{(d)}(\mathbf{x})Y_{k,i}^{(d)}(\sqrt{d}\beta)\\ &= \alpha(\beta^T\mathbf{x})^p,\end{aligned}$$

as desired. To obtain a norm bound on a, we can write

$$\begin{aligned}\|a\|_{L^2}^2 &= \sum_{k=0}^{p}\sum_{i=1}^{B(d,k)}c_{k,i}^2\\ &= \sum_{k=0}^{p}\frac{\lambda_{d,k}(f)^2}{\lambda_{d,k}(\sigma)^2}\sum_{i=1}^{B(d,k)}Y_{k,i}^{(d)}(\beta\sqrt{d})^2\\ &= \sum_{k=0}^{p}\frac{\lambda_{d,k}(f)^2}{\lambda_{d,k}(\sigma)^2}B(d,k)\\ &\lesssim d^p\sum_{k=0}^{p}\lambda_{d,k}(f)^2B(d,k)\\ &= d^p\|f\|_{L^2([-\sqrt{d},\sqrt{d}],\tau_{d-1}^1)}\\ &\lesssim d^p,\end{aligned}$$

since

$$\|f\|_{L^2([-\sqrt{d},\sqrt{d}],\tau_{d-1}^1)} \to_{d\to\infty} \|f\|_{L^2(\mathbb{R},\gamma)},$$

and thus $\|f\|_{L^2([-\sqrt{d},\sqrt{d}],\tau_{d-1}^1)} = \Theta_d(1)$. □

Lemma 5 (Expressivity via infinitely many neurons)**.** *Let $k \geq 1$, and let σ be a twice-differentiable activation such that σ'' satisfies Assumption 3. Then, there exist functions $\mathbf{w}_+, \mathbf{w}_- : \mathcal{S}^{d-1}(1) \to \mathbb{R}$ such that*

$$\mathbb{E}_{\mathbf{w}_0}\left[\sigma''(\mathbf{w}_0^T\mathbf{x})\left((\mathbf{w}_+^T\mathbf{x})^2 - (\mathbf{w}_-^T\mathbf{x})^2\right)\right] = \alpha(\beta^T\mathbf{x})^{k+1},$$

and

$$\mathbb{E}_{\mathbf{w}_0}\left[\|\mathbf{w}_+\|_2^4 + \|\mathbf{w}_-\|_2^4\right] \lesssim d^{k-1}$$

Proof. Note that since σ'' is continuous and bounded, $\sigma'' \in L^2([-\sqrt{d},\sqrt{d}],\tau_{d-1}^1)$. Therefore applying Lemma 4 with activation σ'' and degree $k-1$, there exists a function a satisfying

$$\mathbb{E}_{\mathbf{w}_0\sim\mathcal{S}^{d-1}(1)}\left[\sigma''(\mathbf{w}_0^T\mathbf{x})a(\mathbf{w}_0)\right] = (\beta^T\mathbf{x})^{k-1}$$

and

$$\|a\|_{L^2(\mathcal{S}^{d-1}(1))}^2 \lesssim d^{k-1}.$$

Define

$$\begin{aligned}\mathbf{w}_+(\mathbf{w}_0) &= \sqrt{\max(0, a(\mathbf{w}_0))}\cdot\beta\\ \mathbf{w}_-(\mathbf{w}_0) &= \sqrt{-\min(0, a(\mathbf{w}_0))}\cdot\beta.\end{aligned}$$

Then

$$(\mathbf{w}_+^T\mathbf{x})^2 - (\mathbf{w}_-^T\mathbf{x})^2 = (\beta^T\mathbf{x})^2\left(\max(0,a(\mathbf{w}_0)) + \min(0,a(\mathbf{w}_0))\right) = a(\mathbf{w}_0)(\beta^T\mathbf{x})^2,$$

so

$$\begin{aligned}\mathbb{E}_{\mathbf{w}_0}\left[\sigma''(\mathbf{w}_0^T\mathbf{x})\left((\mathbf{w}_+^T\mathbf{x})^2 - (\mathbf{w}_-^T\mathbf{x})^2\right)\right] &= \mathbb{E}_{\mathbf{w}_0}\left[\sigma''(\mathbf{w}_0^T\mathbf{x})a(\mathbf{w}_0)\right](\beta^T\mathbf{x})^2\\ &= \alpha(\beta^T\mathbf{x})^{k-1}(\beta^T\mathbf{x})^2\\ &= \alpha(\beta^T\mathbf{x})^{k+1}.\end{aligned}$$

Finally, we have the norm bound

$$\begin{aligned}\mathbb{E}_{\mathbf{w}_0}\left[\|\mathbf{w}_+\|_2^4 + \|\mathbf{w}_-\|_2^4\right] &= \mathbb{E}_{\mathbf{w}_0}\left[\max(0, a(\mathbf{w}_0))^2 + \min(0, a(\mathbf{w}_0))^2\right] \\ &= \mathbb{E}_{\mathbf{w}_0}\left[a(\mathbf{w}_0)^2\right] \\ &\lesssim d^{k-1}.\end{aligned}$$

□

B.1.2 Proof of Lemma 1

Proof. We show this Lemma holds with probability $1 - d^{-10}$.

Define $M := \lfloor \frac{m}{2R} \rfloor$. Define for $i \in [R]$, define the subnetwork $f_Q^i(\mathbf{x}, \mathbf{W}_Q)$ by

$$f_Q^i(\mathbf{x}, \mathbf{W}_Q) := \frac{1}{2\sqrt{m}} \sum_{r=(i-1)M+1}^{iM} \sigma''(\mathbf{w}_{0,r}^T \mathbf{x}) \left((\mathbf{x}^T \mathbf{w}_{Q,r})^2 - (\mathbf{x}^T \mathbf{w}_{Q,r+m/2})^2\right).$$

We will now construct $\mathbf{W}_Q \in \mathbb{R}^{d \times m}$. For $RM < r \leq m/2$, set $\mathbf{w}_{Q,r} = \mathbf{w}_{Q,r+m/2} = \mathbf{0}$. As a result, we have that

$$f_Q(\mathbf{x}, \mathbf{W}_Q) = \sum_{i \in [R]} f_Q^i(\mathbf{x}, \mathbf{W}_Q).$$

Our construction will proceed by expressing $\alpha_i(\beta_i^T \mathbf{x})^{k+1}$ with $f_Q^i(\mathbf{x}, \mathbf{W}_Q)$. For fixed $i \in [R]$, let $\mathbf{w}_+^i, \mathbf{w}_-^i$ be from the infinite width construction in Lemma 5, so that

$$\mathbb{E}_{\mathbf{w}_0}\left[\sigma''(\mathbf{w}_0^T \mathbf{x}) \left(({\mathbf{w}_+^i}^T \mathbf{x})^2 - ({\mathbf{w}_-^i}^T \mathbf{x})^2\right)\right] = \alpha_i(\beta_i^T \mathbf{x})^{k+1}.$$

For integers $(i-1)M + 1 \leq r \leq iM$, define

$$(\mathbf{w}_{Q,r}, \mathbf{w}_{Q,r+m/2}) = \sqrt{2/M} \cdot (m^{1/4} \mathbf{w}_+^i(\mathbf{w}_{0,r}), m^{1/4} \mathbf{w}_-^i(\mathbf{w}_{0,r})).$$

Then,

$$\begin{aligned}f_Q^i(\mathbf{x}; \mathbf{W}_Q) &= \frac{1}{2\sqrt{m}} \sum_{r=(i-1)M+1}^{iM} \sigma''(\mathbf{w}_{0,r}^T \mathbf{x}) \left((\mathbf{x}^T \mathbf{w}_{Q,r})^2 - (\mathbf{x}^T \mathbf{w}_{Q,r+m/2})^2\right) \\ &= \frac{1}{M} \sum_{r=(i-1)M+1}^{iM} \sigma''(\mathbf{w}_{0,r}^T \mathbf{x}) a^i(\mathbf{w}_{0,r}) (\beta^T \mathbf{x})^2.\end{aligned}$$

Note that we can bound

$$\begin{aligned}|a^i(\mathbf{w}_0)| &\leq \sum_{k'=0}^{k-1} \sum_{j=1}^{B(d,k')} \left| c_{k',j} Y_{k,j}^{(d)}(\sqrt{d}\mathbf{w}_0) \right| \\ &\leq \left(\sum_{k'=0}^{k-1} \sum_{j=1}^{B(d,k')} c_{k',j}^2 \right)^{1/2} \left(\sum_{k'=0}^{k-1} \sum_{i=1}^{B(d,k')} Y_{k',j}^{(d)}(\sqrt{d}\mathbf{w}_0)^2 \right)^{1/2} \\ &= \|a^i\|_{L^2} \left(\sum_{k'=0}^{k-1} B(d,k) \right)^{1/2} \\ &\lesssim d^{k-1}.\end{aligned}$$

Therefore letting $Z_r = \sigma''(\mathbf{w}_{0,r}^T \mathbf{x}) a^i(\mathbf{w}_{0,r}) (\beta^T \mathbf{x})^2$, we have $|Z_r| \lesssim d^k$. Also, the Z_r are i.i.d and satisfy $\mathbb{E}[Z_r] = (\beta^T \mathbf{x})^{k+1}$. Therefore by Hoeffding's inequality, with probability $1 - \frac{1}{2} d^{-11} n^{-1}$ we have

$$\begin{aligned}\left| f_Q^i(\mathbf{x}_j; \mathbf{W}_Q) - \alpha_i(\beta_i^T \mathbf{x}_j)^{k+1} \right| &= \left| \frac{1}{M} \sum_{r=(i-1)M+1}^{iM} Z_r - \mathbb{E}[Z_r] \right| \\ &= \tilde{O}\left(\frac{d^k}{\sqrt{m}}\right),\end{aligned}$$

where we omit poly(R) dependencies inside the big O notation.

Union bounding over $j \in [n]$, with probability $1 - \frac{1}{2}d^{-11}$ over the initialization,

$$\max_{j\in[n]} \left| f_Q^i(\mathbf{x}_j; \mathbf{W}_Q) - \alpha_i(\beta_i^T \mathbf{x}_j)^{k+1} \right| \leq \tilde{O}\left(\frac{d^k}{\sqrt{m}}\right).$$

Since the above holds for all $i \in [R]$, union bounding over i yields that with probability $1 - \frac{1}{2}d^{-11}R \geq 1 - \frac{1}{2}d^{-10}$,

$$\begin{aligned}\max_{j\in[n]} |f_Q(\mathbf{x}_j, \mathbf{W}_Q) - f_{sp}(\mathbf{x})| &\leq \sum_{i\in[R]} \max_{j\in[n]} \left| f_Q^i(\mathbf{x}_j; \mathbf{W}_Q) - \alpha_i(\beta_i^T \mathbf{x}_j)^{k+1} \right| \\ &\leq \tilde{O}\left(\frac{d^k}{\sqrt{m}}\right).\end{aligned}$$

To bound the norm of $\mathbf{W}_Q$, observe that

$$\|\mathbf{W}_Q\|_{2,4}^4 = \frac{4m}{M^2} \sum_{i\in[R]} \sum_{r=(i-1)M+1}^{iM} \|\mathbf{w}_+^i(\mathbf{w}_{0,r})\|_2^4 + \|\mathbf{w}_-^i(\mathbf{w}_{0,r})\|_2^4.$$

Since we can upper bound

$$\|\mathbf{w}_+^i(\mathbf{w}_{0,r})\|_2^4 + \|\mathbf{w}_-^i(\mathbf{w}_{0,r})\|_2^4 \leq a^i(\mathbf{w}_{0,r})^2 \lesssim d^{2(k-1)},$$

for fixed $i \in [R]$ by Hoeffding we have that with probability $1 - \frac{1}{2}d^{-11}$ over the initialization

$$\left| \frac{1}{M} \sum_{r=(i-1)M+1}^{iM} \|\mathbf{w}_+^i(\mathbf{w}_{0,r})\|_2^4 + \|\mathbf{w}_-^i(\mathbf{w}_{0,r})\|_2^4 - \mathbb{E}_{\mathbf{w}_0}\left[\|\mathbf{w}_+^i\|_2^4 + \|\mathbf{w}_-^i\|_2^4\right] \right| \leq \tilde{O}\left(\frac{d^{2(k-1)}}{\sqrt{m}}\right).$$

Union bounding over each $i \in [R]$ and using $M = \Theta(m/R) = \Theta(m)$, with probability $1 - \frac{1}{2}d^{-10}$ over the initialization we have that

$$\begin{aligned}\|\mathbf{W}_Q\|_{2,4}^4 &\lesssim \sum_{i\in[R]} \mathbb{E}_{\mathbf{w}_0}\left[\|\mathbf{w}_+^i\|_2^4 + \|\mathbf{w}_-^i\|_2^4\right] + \frac{d^{2(k-1)}}{\sqrt{m}} \\ &\lesssim d^{k-1} + \frac{d^{2(k-1)}}{\sqrt{m}} \\ &\lesssim d^{k-1},\end{aligned}$$

as desired. □

Corollary 2. *The solution $\mathbf{W}_Q$ constructed in Lemma 1 satisfies*

$$\|\mathbf{W}_Q\|_{2,\infty} \lesssim m^{-1/4} d^{\frac{k-1}{2}}.$$

Proof. From the proof of Lemma 1, either $\|\mathbf{w}_r\|_2 = 0$, or $(i-1)M + 1 \leq r \leq iM$ or $m/2 + (i-1)M + 1 \leq r \leq m/2 + iM$ for some i in which case

$$\|\mathbf{w}_r\|_2 \leq 2m^{-1/4}\sqrt{|a^i(\mathbf{w}_0)|} \leq 2m^{-1/4} d^{\frac{k-1}{2}}.$$

□

B.2 NTK Proofs

B.2.1 Symmetric Initialization

Recall the definition of the NTK featurization map

$$\varphi(\mathbf{x}) = \text{vec}(\nabla_{\mathbf{W}} f(\mathbf{x}; \mathbf{W}_0)) = \text{vec}(\{\frac{a_r}{\sqrt{m}} \sigma'(\mathbf{w}_{0,r}^T \mathbf{x})\mathbf{x}\}_{r\in[m]}) \in \mathbb{R}^{md}$$

The symmetric initialization makes this different from the NTK features in [25, 39], which for width $\tilde{m} = m/2$ is given by

$$\tilde{\varphi}(\mathbf{x}) = \text{vec}(\nabla_{\mathbf{W}} f(\mathbf{x}; \mathbf{W}_0)) = \text{vec}(\{\frac{1}{\sqrt{m}}\sigma'(\mathbf{w}_{0,r}^T \mathbf{x})\mathbf{x}\}_{r\in[\tilde{m}]}) \in \mathbb{R}^{\tilde{m}d}$$

These two features are related by

$$\varphi(\mathbf{x}) = \begin{bmatrix} \tilde{\varphi}(\mathbf{x}) \\ -\tilde{\varphi}(\mathbf{x}) \end{bmatrix}.$$

For the bulk of this section we consider the features $\tilde{\varphi}(\mathbf{x})$ in order to invoke the results from [25, 39].

B.2.2 Preliminaries

For arbitary $N \ll md$, let $\mathcal{D}_N = \{\mathbf{x}_i\}_{i\in[N]}$ be a dummy dataset of size N, where each $\mathbf{x}_i$ is sampled i.i.d from $\mathcal{S}^{d-1}(\sqrt{d})$. We define the following random matrices which depend on $\mathcal{D}_N$.

Denote by

$$\mathbf{\Phi}_N = \begin{bmatrix} \tilde{\varphi}(\mathbf{x}_1)^T \\ \tilde{\varphi}(\mathbf{x}_2)^T \\ \cdots \\ \tilde{\varphi}(\mathbf{x}_N)^T \end{bmatrix} \in \mathbb{R}^{N\times \tilde{m}d}$$

the feature matrix, and let

$$\mathbf{K}_N = \mathbf{\Phi}_N \mathbf{\Phi}_N^T \in \mathbb{R}^{N\times N}$$

be the empirical kernel matrix.

The infinite-width kernel matrix $\mathbf{K}_N^\infty \in \mathbb{R}^{N\times N}$ has entries

$$\{\mathbf{K}_N^\infty\}_{i,j} = \mathbb{E}_{\mathbf{w}}[\sigma'(\mathbf{w}^T\mathbf{x}_i)\sigma'(\mathbf{w}^T\mathbf{x}_j)\mathbf{x}_i^T\mathbf{x}_j].$$

Also, define $\mathbf{\Sigma}_N \in \mathbb{R}^{\tilde{m}d\times\tilde{m}d}$ to be the empirical covariance matrix, so that

$$\mathbf{\Sigma}_N = \frac{1}{N}\mathbf{\Phi}_N^T\mathbf{\Phi}_N = \frac{1}{N}\sum_{i=1}^N \tilde{\varphi}(\mathbf{x}_i)\tilde{\varphi}(\mathbf{x}_i)^T,$$

and let

$$\tilde{\mathbf{\Sigma}} = \mathbb{E}_\mu\left[\tilde{\varphi}(\mathbf{x})\tilde{\varphi}(\mathbf{x})^T\right] \in \mathbb{R}^{\tilde{m}d\times\tilde{m}d}$$

be the population covariance matrix.

We let σ' satisfy assumption 3. Along with the boundedness of σ', this allows us to invoke the following lemmas from [39]:

Lemma 6 ([39], Theorem 3.2)**.** *With probability* $1 - d^{-11}$,

$$\|\{\mathbf{K}_N^\infty\}^{-\frac{1}{2}}\mathbf{K}_N\mathbf{K}_N^{\infty\,-\frac{1}{2}} - \mathbf{I}_N\|_{op} \le \tilde{O}\left(\sqrt{\frac{N}{\tilde{m}d}} + \frac{N}{\tilde{m}d}\right)$$

As in [39], let $\mathbf{\Psi}_{\le\ell} \in \mathbb{R}^{N\times n_\ell}$ be the evaluations of the degree $\le \ell$ spherical harmonics on the N data points.

Lemma 7 ([39], Lemma 2)**.** *Let* $d^\ell \log^2 d \ll N \ll d^{\ell+1}/\log d^C$. *With probability* $1 - d^{-11}$ *the infinite-width kernel matrix can be decomposed as*

$$\frac{1}{d}\mathbf{K}_N^\infty := \gamma_{>\ell}\mathbf{I}_N + \mathbf{\Psi}_{\le\ell}\mathbf{\Lambda}_{\le\ell}^2\mathbf{\Psi}_{\le\ell}^T + \mathbf{\Delta},$$

where $\gamma_k \ge 0$ *is a sequence satisfying*

$$\gamma_0 = d^{-1}(1 + o_d(1))\mu_1^2(\sigma'), \quad \gamma_k = \mu_{k-1}^2(\sigma') + o_d(1) \text{ for } k \ge 1, \quad \gamma_{>\ell} := \sum_{k'>\ell}\gamma_k \tag{43}$$

and $\mathbf{\Lambda}^2_{\le \ell}$ is a diagonal matrix where $B(d,k)^{-1}\gamma_k$ has multiplicity $B(d,k)$, for $k \le \ell$. Furthermore, the remainder $\mathbf{\Delta}$ satisfies

$$\|\mathbf{\Delta}\|_{op} \le \tilde{O}\left(\sqrt{\frac{N}{d^{\ell+1}}}\right),$$

and the spherical harmonic features $\mathbf{\Psi}_{\le \ell}$ satisfy

$$\left\|\frac{1}{N}\mathbf{\Psi}_{\le \ell}^T\mathbf{\Psi}_{\le \ell} - \mathbf{I}_{n_\ell}\right\|_{op} \le \tilde{O}\left(\sqrt{\frac{d^\ell}{N}}\right).$$

By assumption 3, $\gamma_0 = \Theta(d^{-1})$ and $\gamma_{\ell'} = \Theta(1)$ for $\ell' \ge 1$.

The following Lemma gives the eigenvalues of the empirical kernel matrix $\mathbf{K}_N$ and the infinite-width kernel matrix $\mathbf{K}_N^\infty$:

Lemma 8 (Follows from [39], Lemma 6)**.** *With probability $1 - d^{-11}$, the following all hold:*

For $1 \le \ell' \le \ell$, $n_{\ell'-1} < i \le n_{\ell'}$,

$$\lambda_i(\mathbf{K}_N^\infty), \lambda_i(\mathbf{K}_N) = \Theta(N \cdot d^{1-\ell'}).$$

Additionally,

$$\lambda_1(\mathbf{K}_N^\infty), \lambda_1(\mathbf{K}_N) = \Theta(N).$$

Finally, for $i > n_\ell$,

$$\lambda_i(\mathbf{K}_N^\infty), \lambda_i(\mathbf{K}_N) = \Theta(d).$$

We also use the following classical results throughout the proofs in this section.

Lemma 9 (Weyl's Inequality)**.** *For two psd matrices $\mathbf{A}_1, \mathbf{A}_2 \in \mathbb{R}^{p\times p}$,*

$$|\lambda_i(\mathbf{A}_1) - \lambda_i(\mathbf{A}_2)| \le \|\mathbf{A}_1 - \mathbf{A}_2\|_{op} \tag{44}$$

for all $i \in [p]$.

Definition 3 ([13])**.** *For $r \le p$, let $\mathbf{U}_1, \mathbf{U}_2 \in \mathbb{R}^{p\times r}$ both have orthonormal columns. Then the distance between the subspaces spanned by $\mathbf{U}_1, \mathbf{U}_2$ is*

$$\mathrm{dist}(\mathbf{U}_1, \mathbf{U}_2) := \min_{\mathbf{O}\in\mathcal{O}^{r\times r}} \|\mathbf{U}_1\mathbf{O} - \mathbf{U}_2\|_{op}, \tag{45}$$

where $\mathcal{O}^{r\times r}$ is the space of $r \times r$ orthogonal matrices.

Lemma 10 (Davis-Kahan sin-θ theorem [17, 13])**.** *For two psd matrices $\mathbf{A}_1, \mathbf{A}_2 \in \mathbb{R}^{p\times p}$, let $\mathbf{A}_1 = \mathbf{U}_1\mathbf{\Lambda}_1\mathbf{U}_1^T$ and $\mathbf{A}_2 = \mathbf{U}_2\mathbf{\Lambda}_2\mathbf{U}_2^T$ be their eigendecompositions (sorted by decreasing eigenvalues), and let $\mathbf{U}_1 = [\mathbf{U}_{1,\le r} \quad \mathbf{U}_{1,>r}]$, $\mathbf{U}_2 = [\mathbf{U}_{2,\le r} \quad \mathbf{U}_{2,>r}]$ be their eigenvectors, where $\mathbf{U}_{1,\le r}, \mathbf{U}_{2,\le r} \in \mathbb{R}^{p\times r}$. Furthermore, assume $\|\mathbf{A}_1 - \mathbf{A}_2\|_{op} < (1 - 1/\sqrt{2})(\lambda_r(\mathbf{A}_1) - \lambda_{r+1}(\mathbf{A}_1))$. Then*

$$\mathrm{dist}(\mathbf{U}_{1,\le r}, \mathbf{U}_{2,\le r}) \le \frac{2\|\mathbf{A}_1 - \mathbf{A}_2\|_{op}}{\lambda_r(\mathbf{A}_1) - \lambda_{r+1}(\mathbf{A}_1)} \tag{46}$$

B.2.3 Eigenvector Lemmas

Throughout this section, we condition on the event where Lemmas 6, 7, 8 are all true.

Lemma 11. *Define*

$$\tilde{\mathbf{K}} = d\gamma_{>k}\mathbf{I}_N + d\mathbf{\Psi}_{\le \ell}\mathbf{\Lambda}^2_{\le \ell}\mathbf{\Psi}^T_{\le \ell},$$

and let $\tilde{\mathbf{K}}$ have eigendecomposition $\tilde{\mathbf{U}}\tilde{\mathbf{\Lambda}}^2\tilde{\mathbf{U}}^T$. Define $\mathbf{\Psi}_{\le k} \in \mathbb{R}^{N\times n_k}$ to be the first n_k columns of $\mathbf{\Psi}_{\le k}$ (where $k < \ell$), and also let $\tilde{\mathbf{U}}_{\le k}$ be the first n_k columns of $\tilde{\mathbf{U}}$, with $\tilde{\mathbf{U}}_{>k}$ the remaining columns. Then

$$\left\|\frac{1}{\sqrt{N}}\mathbf{\Psi}^T_{\le k}\tilde{\mathbf{U}}_{>k}\right\|_{op} \lesssim \frac{1}{\sqrt{d}}.$$

Proof. Let $\mathbf{\Psi}_{k:\ell} \in \mathbb{R}^{N \times n_\ell - n_k}$ be the (n_k+1)-th to n_ℓ-th columns. We can then decompose

$$\tilde{\mathbf{K}} = d\mathbf{\Psi}_{\le k}\mathbf{\Lambda}_{\le k}^2\mathbf{\Psi}_{\le k}^T + d\mathbf{\Psi}_{k:\ell}\mathbf{\Lambda}_{k:\ell}^2\mathbf{\Psi}_{k:\ell}^T + d\gamma_{>k}\mathbf{I}_N$$

and

$$\tilde{\mathbf{K}} = \begin{bmatrix}\tilde{\mathbf{U}}_{\le k} & \tilde{\mathbf{U}}_{>k}\end{bmatrix} \operatorname{diag}(\tilde{\mathbf{\Lambda}}_{\le k}, \tilde{\mathbf{\Lambda}}_{>k}) \begin{bmatrix}\tilde{\mathbf{U}}_{\le k} & \tilde{\mathbf{U}}_{>k}\end{bmatrix}^T,$$

where $\tilde{\mathbf{U}}_{\le k} \in \mathbb{R}^{N \times n_k}$. Let $\mathbf{u} \in \mathbb{R}^{N-n_k}$ be a vector such that $\|\mathbf{u}\|_2 = 1$ and $\|\mathbf{\Psi}_{\le k}^T\tilde{\mathbf{U}}_{>k}\mathbf{u}\|_2 = \|\mathbf{\Psi}_{\le k}^T\tilde{\mathbf{U}}_{>k}\|_{op}$. Finally, define $\tilde{\mathbf{u}} = \mathbf{\Psi}_{\le k}^T\tilde{\mathbf{U}}_{>k}\mathbf{u}$. We then have:

$$\begin{aligned}
\mathbf{u}^T\tilde{\mathbf{\Lambda}}_{>k}\mathbf{u} &= \mathbf{u}^T\tilde{\mathbf{U}}_{>k}^T\tilde{\mathbf{K}}\tilde{\mathbf{U}}_{>k}\mathbf{u} \\
&\ge d\mathbf{u}^T\tilde{\mathbf{U}}_{>k}^T\mathbf{\Psi}_{\le k}\mathbf{\Lambda}_{\le k}^2\mathbf{\Psi}_{\le k}^T\tilde{\mathbf{U}}_{>k}\mathbf{u} + d\gamma_{>k}. \\
&= d\tilde{\mathbf{u}}^T\mathbf{\Lambda}_{\le k}^2\tilde{\mathbf{u}} + d\gamma_{>k} \\
&\gtrsim d \cdot d^{-k}\|\tilde{\mathbf{u}}\|_2^2 + d\gamma_{>k},
\end{aligned}$$

since $\lambda_{min}(\mathbf{\Lambda}_{\le k}^2) = \Theta(d^{-k})$.

Define

$$\tilde{\mathbf{K}}_{\le k} = d\mathbf{\Psi}_{\le k}\mathbf{\Lambda}_{\le k}^2\mathbf{\Psi}_{\le k}^T,$$

and define $\tilde{\mathbf{K}}_{>k} = \tilde{\mathbf{K}} - \tilde{\mathbf{K}}_{\le k}$. By Weyl's inequality,

$$\left|\lambda_i(\mathbf{K}_{\le k}) - \lambda_i(\tilde{\mathbf{K}})\right| \le \|\tilde{\mathbf{K}}_{>k}\|_{op}.$$

For $i > n_k$, $\lambda_i(\mathbf{K}_{\le k}) = 0$, in which case

$$\lambda_i(\tilde{\mathbf{K}}) \le \|\tilde{\mathbf{K}}_{>k}\|_{op} \lesssim dN \cdot d^{-k-1}.$$

Therefore we can upper bound

$$\mathbf{u}^T\tilde{\mathbf{\Lambda}}_{>k}\mathbf{u} \lesssim dN \cdot d^{-k-1}.$$

Altogether, we have

$$dN \cdot d^{-k-1} \gtrsim d \cdot d^{-k}\|\tilde{\mathbf{u}}\|_2^2 + d\gamma_{>k},$$

which yields

$$\frac{1}{\sqrt{N}}\|\tilde{\mathbf{u}}\|_2 \lesssim \frac{1}{\sqrt{d}}.$$

By definition, $\|\tilde{\mathbf{u}}\|_2 = \|\mathbf{\Psi}_{\le k}^T\tilde{\mathbf{U}}_{>k}\|_{op}$, so

$$\left\|\frac{1}{\sqrt{N}}\mathbf{\Psi}_{\le k}^T\tilde{\mathbf{U}}_{>k}\right\|_{op} \lesssim \frac{1}{\sqrt{d}},$$

as desired. □

Lemma 12. *Let $\mathbf{K}_N^\infty$ have eigendecomposition $\mathbf{K}_N^\infty = \overline{\mathbf{U}}\overline{\mathbf{\Lambda}}^2\overline{\mathbf{U}}^T$, where $\overline{\mathbf{U}}_{\le k}, \overline{\mathbf{U}}_{>k}$ are the first n_k, remaining $N - n_k$ columns of $\overline{\mathbf{U}}$ respectively. Then*

$$\operatorname{dist}\left(\tilde{\mathbf{U}}_{\le k}, \overline{\mathbf{U}}_{\le k}\right) \le \tilde{O}\left(\sqrt{\frac{d^{2k-\ell-1}}{N}}\right)$$

Proof. By Lemma 8, $\lambda_{n_k}(\mathbf{K}_N^\infty) - \lambda_{n_k+1}(\mathbf{K}_N^\infty) \gtrsim dN \cdot d^{-k}$. Since

$$\|\mathbf{K}_N^\infty - \tilde{\mathbf{K}}\|_{op} = d\|\mathbf{\Delta}\|_{op} \le \tilde{O}\left(\sqrt{\frac{N}{d^{\ell-1}}}\right) \ll \lambda_{n_k}(\mathbf{K}_N^\infty) - \lambda_{n_k+1}(\mathbf{K}_N^\infty)$$

by Davis-Kahan we have

$$\operatorname{dist}\left(\tilde{\mathbf{U}}_{\le k}, \overline{\mathbf{U}}_{\le k}\right) \lesssim \frac{d\|\mathbf{\Delta}\|_{op}}{dN \cdot d^{-k}} = \tilde{O}\left(\sqrt{\frac{d^{2k-\ell-1}}{N}}\right).$$

□

Lemma 13. *Let $\mathbf{K}_N$ have eigendecomposition $\mathbf{K}_N = \mathbf{U}_N \mathbf{\Lambda}_N^2 \mathbf{U}_N^T$. Then*

$$\text{dist}\left(\overline{\mathbf{U}}_{\leq k}, \mathbf{U}_{N,\leq k}\right) = \tilde{O}\left(\sqrt{\frac{Nd^{2k-3}}{\tilde{m}}}\right)$$

Proof. By Lemma 6, we can write

$$\mathbf{K}_N^{\infty -1/2} \mathbf{K}_N \mathbf{K}_N^{\infty -1/2} = \mathbf{I}_N + \mathbf{\Delta}_N,$$

where $\mathbf{\Delta}_N \leq \tilde{O}\left(\sqrt{\frac{N}{\tilde{m}d}}\right)$. Rearranging, we get

$$\mathbf{K}_N = \mathbf{K}_N^{\infty} + \mathbf{K}_N^{\infty 1/2} \mathbf{\Delta}_N \mathbf{K}_N^{\infty 1/2},$$

where we can bound

$$\|\mathbf{K}_N^{\infty 1/2} \mathbf{\Delta}_N \mathbf{K}_N^{\infty 1/2}\|_{op} \leq \|\mathbf{\Delta}_N\|_{op} \|\mathbf{K}_N^{\infty}\|_{op} \lesssim \sqrt{\frac{N^{5/2}}{\tilde{m}d}},$$

since $\|\mathbf{K}_N^{\infty}\|_{op} \lesssim N$. Since $\mathbf{K}_N^{\infty}$ has eigengap $\Theta(N \cdot d^{1-k})$ we can again apply Davis-Kahan to get

$$\text{dist}\left(\overline{\mathbf{U}}_{\leq k}, \mathbf{U}_{N,\leq k}\right) \lesssim \frac{\|\mathbf{K}_N^{\infty 1/2} \mathbf{\Delta}_n \mathbf{K}_N^{\infty 1/2}\|_{op}}{N \cdot d^{1-k}} = \tilde{O}\left(\sqrt{\frac{Nd^{2k-3}}{\tilde{m}}}\right).$$

□

Lemma 14. *Let $\mathbf{\Sigma}_N$ have eigendecomposition $\mathbf{\Sigma}_N = \mathbf{V}_N \frac{\mathbf{\Lambda}_N^2}{N} \mathbf{V}_N^T$, where $\mathbf{V}_N \in \mathbb{R}^{md \times N}$. Let $\tilde{\mathbf{\Sigma}}$ have eigendecomposition $\tilde{\mathbf{\Sigma}} = \mathbf{V} \mathbf{\Lambda}^2 \mathbf{V}^T$. Let $\mathbf{V}_N = [\mathbf{V}_{N,\leq k} \quad \mathbf{V}_{N,>k}]$, $\mathbf{V} = [\mathbf{V}_{\leq k} \quad \mathbf{V}_{>k}]$ where $\mathbf{V}_{N,\leq k}, \mathbf{V}_{\leq k} \in \mathbb{R}^{md \times n_k}$. Then, with probability $1 - n^{-11}$,*

$$\text{dist}\left(\mathbf{V}_{N,\leq k}, \mathbf{V}_{\leq k}\right) \leq \tilde{O}\left(\sqrt{\frac{d^{2k}}{N}}\right).$$

Proof. By [42, 5.6.4], with probability $1 - n^{-11}$ we have

$$\|\mathbf{\Sigma}_N - \tilde{\mathbf{\Sigma}}\|_{op} \leq \tilde{O}\left(\sqrt{\frac{r}{N}}\right) \cdot \|\tilde{\mathbf{\Sigma}}\|_{op},$$

where $r = \text{Tr}(\tilde{\mathbf{\Sigma}})/\|\tilde{\mathbf{\Sigma}}\|_{op}$ is the effective rank. We can upper bound $\text{Tr}(\tilde{\mathbf{\Sigma}}) \leq d$, and thus

$$\|\mathbf{\Sigma}_N - \tilde{\mathbf{\Sigma}}\|_{op} \leq \tilde{O}\left(\sqrt{\frac{d}{N}\|\tilde{\mathbf{\Sigma}}\|_{op}}\right) = \tilde{O}\left(\sqrt{\frac{d^2}{N}}\right),$$

since $\|\tilde{\mathbf{\Sigma}}\|_{op} \leq d$. By Lemma 8,

$$\lambda_{n_k}(\mathbf{\Sigma}_N) - \lambda_{n_k+1}(\mathbf{\Sigma}_N) = \frac{1}{N} \cdot \Theta(N \cdot d^{1-k}) = \Theta(d^{1-k}).$$

Therefore by Davis-Kahan, we can bound

$$\text{dist}\left(\mathbf{V}_{N,\leq k}, \mathbf{V}_{\leq k}\right) \lesssim \cdot \frac{\|\mathbf{\Sigma}_N - \tilde{\mathbf{\Sigma}}\|_{op}}{d^{-k+1}} \lesssim \sqrt{\frac{d^{2k}}{N}}.$$

□

The following lemma is a consequence of the preceeding eigenstructure and matrix perturbation lemmas, and partitions the eigenvectors of $\mathbf{\Sigma}$ into groups corresponding to large, medium, and small eigenvalues.

Lemma 15. *Let $N \geq d^{4k}, m \geq N^{5/2}$. For $1 \leq k' \leq 2k$, $n_{k'-1} < i \leq n_{k'}$, we have $\lambda_i(\mathbf{\Sigma}) = \Theta(d^{1-k'})$. Also, for $i > n_{2k}$, $\lambda_i(\mathbf{\Sigma}) \ll \lambda_{n_{2k}}(\mathbf{\Sigma})$.*

Proof. Since

$$\|\mathbf{\Sigma}_N - \tilde{\mathbf{\Sigma}}\|_{op} \lesssim \sqrt{\frac{d^2}{N}},$$

by Weyl's inequality we have

$$|\lambda_i(\mathbf{\Sigma}_N) - \lambda_i(\tilde{\mathbf{\Sigma}})| \lesssim \sqrt{\frac{d^2}{N}}$$

Since for $n_{k'-1} < i \leq n_{k'}$ we have $\lambda_i(\mathbf{\Sigma}_N) = N^{-1}\lambda_i(\mathbf{K}_N) = \Theta(d^{1-k'}) \gg \sqrt{\frac{d^2}{N}}$, since $N \geq d^{4k}$; therefore $\lambda_i(\mathbf{\Sigma}) = \Theta(d^{1-k'})$. Furthermore, for $i > n_{2k}$,

$$\lambda_i(\mathbf{\Sigma}) \lesssim \sqrt{\frac{d^2}{N}} + \lambda_i(\mathbf{\Sigma}_N) \ll \Theta(d^{1-2k}).$$

As a consequence, we can write the eigendecomposition of $\mathbf{\Sigma}$ as

$$\mathbf{\Sigma} = [\mathbf{Q}_1 \quad \mathbf{Q}_2 \quad \mathbf{Q}_3] \begin{bmatrix} \mathbf{\Lambda}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Lambda}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{\Lambda}_3 \end{bmatrix} \begin{bmatrix} \mathbf{Q}_1^T \\ \mathbf{Q}_2^T \\ \mathbf{Q}_3^T \end{bmatrix},$$

where $\mathbf{Q}_1 \in \mathbb{R}^{md \times n_k}$ are the large eigenvectors, $\mathbf{Q}_2 \in \mathbb{R}^{md \times n_{2k} - n_k}$ are the medium eigenvectors, and $\mathbf{Q}_3 \in \mathbb{R}^{md \times md - n_{2k}}$ are the small eigenvectors; concretely,

$$\lambda_{min}(\mathbf{\Lambda}_1) = \Theta(d^{1-k}) \gg \Theta(d^{-k}) = \lambda_{max}(\mathbf{\Lambda}_2)$$

and

$$\lambda_{min}(\mathbf{\Lambda}_2) = \Theta(d^{1-2k}) \gg \lambda_{max}(\mathbf{\Lambda}_3).$$

□

B.2.4 Proof of Lemma 2

Proof. We show this Lemma holds with probability $1 - d^{-10}$. Condition on the events where Lemmas 6, 7, 8, 14 hold.

Pick N so that $d^\ell \ll N \ll d^{\ell+1}$ for $\ell = 4k$, and choose $\tilde{m} = N^{5/2}$. We form a dataset of N samples by adding another $N - n$ samples i.i.d from $\mathcal{S}^{d-1}(\sqrt{d})$. Let $\mathbf{y}_N \in \mathbb{R}^N$ be the vector where $\mathbf{y}_{N,i} = f_k(\mathbf{x}_i)$ where $i \in [N]$. Recall that $\mathbf{\Psi}_{\leq k} \in \mathbb{R}^{N \times n_k}$ denotes the evaluations of the degree $\leq k$ spherical harmonics on the N data points. Since $f_k(\mathbf{x})$ is a degree $\leq k$ polynomial, we can orthogonally decompose it in terms of the degree $\leq k$ spherical harmonics, i.e

$$f_k(\mathbf{x}) = \sum_{k'=0}^{k} \sum_{t=1}^{B(d,k)} w_{k't} Y_{k't}(\mathbf{x}),$$

where $\sum_{k'=0}^{k} \sum_{t=1}^{B(d,k)} w_{k't}^2 = 1$. Therefore we have

$$\mathbf{y}_N = \mathbf{\Psi}_{\leq k} \mathbf{w}^*$$

for $\mathbf{w}^* \in \mathbb{R}^{n_k}$ where $\|\mathbf{w}^*\|_2 = 1$.

Observe that $\mathbf{\Phi}_N$ has SVD $\mathbf{\Phi}_N = \mathbf{U}_N \mathbf{\Lambda}_N \mathbf{V}_N^T$. Define $\tilde{\mathbf{w}}^* = \tilde{\mathbf{U}}_{\leq k}^T \mathbf{\Psi}_{\leq k} \mathbf{w}^*$. Then

$$\begin{aligned} \tilde{\mathbf{U}}_{\leq k} \tilde{\mathbf{w}}^* &= \tilde{\mathbf{U}}_{\leq k} \tilde{\mathbf{U}}_{\leq k}^T \mathbf{\Psi}_{\leq k} \mathbf{w}^* \\ &= (\mathbf{I}_{n_k} - \tilde{\mathbf{U}}_{>k} \tilde{\mathbf{U}}_{>k}^T) \mathbf{\Psi}_{\leq k} \mathbf{w}^* \\ &= \mathbf{y}_N - \tilde{\mathbf{U}}_{>k} \tilde{\mathbf{U}}_{>k}^T \mathbf{\Psi}_{\leq k} \mathbf{w}^*, \end{aligned}$$

and thus

$$\begin{aligned} \|\mathbf{y}_N - \tilde{\mathbf{U}}_{\leq k} \tilde{\mathbf{w}}^*\| &\leq \|\tilde{\mathbf{U}}_{>k} \tilde{\mathbf{U}}_{>k}^T \mathbf{\Psi}_{\leq k} \mathbf{w}^*\| \\ &\leq \|\tilde{\mathbf{U}}_{>k}^T \mathbf{\Psi}_{\leq k}\|_{op} \|\mathbf{w}^*\| \\ &\lesssim \sqrt{\frac{N}{d}}, \end{aligned}$$

by Lemma 11. Also,

$$\|\tilde{\mathbf{w}}^*\| \leq \|\mathbf{\Psi}_{\leq k}\|_{op}\|\mathbf{w}^*\| \lesssim \sqrt{N}.$$

By Lemmas 12 and 13, we have

$$\text{dist}\left(\tilde{\mathbf{U}}_{\leq k}, \mathbf{U}_{N,\leq k}\right) \lesssim \frac{1}{\sqrt{d}}.$$

Thus there exists an orthogonal $\mathbf{O}_1$ such that $\left\|\tilde{\mathbf{U}}_{\leq k} - \mathbf{U}_{N,\leq k}\mathbf{O}_1\right\|_{op} = \text{dist}\left(\tilde{\mathbf{U}}_{\leq k}, \mathbf{U}_{N,\leq k}\right)$ and hence

$$\begin{aligned}\|\mathbf{y}_N - \mathbf{U}_{N,\leq k}\mathbf{O}_1\tilde{\mathbf{w}}^*\| &\leq \|\mathbf{y}_N - \tilde{\mathbf{U}}_{\leq k}\tilde{\mathbf{w}}^*\| + \left\|\tilde{\mathbf{U}}_{\leq k} - \mathbf{U}_{N,\leq k}\mathbf{O}_1\right\|_{op}\|\tilde{\mathbf{w}}^*\| \\ &\lesssim \sqrt{\frac{N}{d}}.\end{aligned}$$

Since $\mathbf{U}_{N,\leq k} = \mathbf{\Phi}_N\mathbf{V}_{N,\leq k}\mathbf{\Lambda}_{N,\leq k}^{-1}$, we have

$$\|\mathbf{y}_N - \mathbf{\Phi}_N\mathbf{V}_{N,\leq k}\mathbf{\Lambda}_{N,\leq k}^{-1}\mathbf{O}_1\tilde{\mathbf{w}}^*\| \lesssim \sqrt{\frac{N}{d}}$$

Finally, by Lemma 14, we have

$$\text{dist}(\mathbf{V}_{N,\leq k}, \mathbf{V}_{\leq k}) \lesssim \sqrt{\frac{d^{2k}}{N}} \leq d^{-k},$$

and thus there exists an orthogonal $\mathbf{O}_2$ such that $\|\mathbf{V}_{N,\leq k} - \mathbf{V}_{\leq k}\mathbf{O}_2\|_{op} \leq d^{-k}$ and hence

$$\begin{aligned}\|\mathbf{y}_N - \mathbf{\Phi}_N\mathbf{V}_{\leq k}\mathbf{O}_2\mathbf{\Lambda}_{N,\leq k}^{-1}\mathbf{O}_1\tilde{\mathbf{w}}^*\| &\leq \|\mathbf{y}_N - \mathbf{\Phi}_N\mathbf{V}_{N,\leq k}\mathbf{\Lambda}_{N,\leq k}^{-1}\mathbf{O}_1\tilde{\mathbf{w}}^*\| + \left\|\mathbf{\Phi}_N(\mathbf{V}_{N,\leq k} - \mathbf{V}_{\leq k}\mathbf{O}_2)\mathbf{\Lambda}_{N,\leq k}^{-1}\mathbf{O}_1\tilde{\mathbf{w}}^*\right\| \\ &\lesssim \sqrt{\frac{N}{d}} + \|\mathbf{\Phi}_N\|_{op}\|\mathbf{V}_{N,\leq k} - \mathbf{V}_{\leq k}\mathbf{O}_2\|_{op}\|\mathbf{\Lambda}_{N,\leq k}^{-1}\|_{op}\|\tilde{\mathbf{w}}^*\| \\ &\lesssim \sqrt{\frac{N}{d}} + \sqrt{N}\cdot d^{-k}\cdot\sqrt{N^{-1}d^{k-1}}\cdot\sqrt{N} \\ &\lesssim \sqrt{\frac{N}{d}}.\end{aligned}$$

Therefore, letting $\mathbf{v}^* = \mathbf{V}_{\leq k}\mathbf{O}_2\mathbf{\Lambda}_{N,\leq k}^{-1}\mathbf{O}_1\tilde{\mathbf{w}}^*$

$$\frac{1}{N}\|\mathbf{y}_N - \mathbf{\Phi}_N\mathbf{v}^*\|_2^2 \lesssim \frac{1}{d}.$$

Note that $\mathbf{v}^*$ satisfies

$$\|\mathbf{v}^*\|_2^2 \leq \|\mathbf{\Lambda}_{N,\leq k}^{-1}\|_{op}^2\|\bar{\mathbf{w}}^*\|^2 = O\left(d^{k-1}\right).$$

Since the $\{\mathbf{x}_i\}_{i\in[N]}$ are i.i.d, we can treat the dataset $\mathcal{D}$ as a subsample of n data points. Since $|f_k(\mathbf{x})|^2 \leq n_k = \Theta(d^k)$, and $|\tilde{\varphi}(\mathbf{x})^T\mathbf{v}^*|^2 \leq \|\tilde{\varphi}(\mathbf{x})\|^2\|\mathbf{v}^*\|^2 = \Theta(d^k)$, we can bound $(f_k(\mathbf{x}) - \tilde{\varphi}(\mathbf{x})^T\mathbf{v}^*)^2 \leq d^k$. Also,

$$\mathbb{E}_N\left[(f_k(\mathbf{x}) - \tilde{\varphi}(\mathbf{x})^T\mathbf{v}^*)^4\right] \lesssim d^k\mathbb{E}_N\left[(f_k(\mathbf{x}) - \tilde{\varphi}(\mathbf{x})^T\mathbf{v}^*)^2\right] \lesssim d^{k-1}.$$

Therefore by Bernstein's Inequality, with probability $1 - d^{-11}$, we can bound

$$\mathbb{E}_n\left[(f_k(\mathbf{x}) - \tilde{\varphi}(\mathbf{x})^T\mathbf{v}^*)^2\right] \lesssim \frac{1}{d} + \sqrt{\frac{d^{k-1}}{n}} + \frac{d^k}{n} \lesssim \frac{d^k}{n}.$$

To conclude we must relate $\tilde{\varphi}$ to φ. Observe that

$$\mathbf{\Sigma} = \begin{bmatrix}\tilde{\mathbf{\Sigma}} & -\tilde{\mathbf{\Sigma}} \\ -\tilde{\mathbf{\Sigma}} & \tilde{\mathbf{\Sigma}}\end{bmatrix}.$$

Therefore $\lambda_i(\mathbf{\Sigma}) = 2\lambda_i(\tilde{\mathbf{\Sigma}})$ for $i \le \tilde{m}$, and $\lambda_i(\mathbf{\Sigma}) = 0$ otherwise. Furthermore, if $\mathbf{v}$ is an eigenvector of $\tilde{\mathbf{\Sigma}}$, then $\begin{bmatrix}\mathbf{v}\\-\mathbf{v}\end{bmatrix}$ is an eigenvector of $\mathbf{\Sigma}$. As a result, if $\mathbf{u} \in \text{span}(\{\mathbf{v}_i(\tilde{\mathbf{\Sigma}})\}_{i\in[n_k]})$, then $\begin{bmatrix}\mathbf{u}\\-\mathbf{u}\end{bmatrix} \in \text{span}(\mathbf{P}_{\le k})$. Letting $\mathbf{z}^* = \frac{1}{2}\begin{bmatrix}\mathbf{v}^*\\-\mathbf{v}^*\end{bmatrix}$, we get that $\mathbf{z}^* \in \text{span}(\mathbf{P}_{\le k})$, and also $\varphi(\mathbf{x})^T\mathbf{z}^* = \tilde{\varphi}(\mathbf{x})^T\mathbf{v}^*$. Therefore

$$\mathbb{E}_n\left[(f_k(\mathbf{x}) - \varphi(\mathbf{x})^T\mathbf{z}^*)^2\right] \lesssim \frac{d^k}{n},$$

and also $\|\mathbf{z}^*\|_2^2 \lesssim d^{k-1}$. Union bounding over all high probability events, this holds with probability $1 - 4d^{-11} \ge 1 - d^{-10}$, as desired. □

B.3 Proof of Theorem 2

Proof. We condition on the events of Lemmas 1, 2, 16, 17, 18 holding, which occurs with probability $1 - 5d^{-10} \ge 1 - d^{-9}$.

We proceed by the probabilistic method. For $r \in [m]$, let the σ_r be random variables with $\sigma_r \sim \text{Unif}(\{\pm 1\})$ i.i.d, and let $\mathbf{S} = \text{diag}(\sigma_1, \dots, \sigma_m)$ be the diagonal matrix of random signs. Let $\mathbf{W}_\mathbf{S}^* = \mathbf{W}_L + \mathbf{W}_Q\mathbf{S}$ be a (random) solution. We will show that in expectation over $\mathbf{S}$ the training error and all the regularizers are small, which implies the existence of a $\mathbf{S}$ which makes all these quantities small.

Bounding the Empirical Loss. First, observe that we can write

$$\begin{aligned}
&\mathbb{E}_n\left|f_L(\mathbf{x};\mathbf{W}_\mathbf{S}^*) + f_Q(\mathbf{x};\mathbf{W}_\mathbf{S}^*) - f_L(\mathbf{x};\mathbf{W}_L) - f_Q(\mathbf{x};\mathbf{W}_Q)\right| &(47)\\
\le\ &\mathbb{E}_n\left|f_L(\mathbf{x};\mathbf{W}_Q\mathbf{S})\right| + \mathbb{E}_\mathcal{D}\left|f_Q(\mathbf{x};\mathbf{W}_S^*) - f_Q(\mathbf{x};\mathbf{W}_Q)\right|. &(48)
\end{aligned}$$

Since $f_Q(\mathbf{x};\mathbf{W}_Q) = f_Q(\mathbf{x};\mathbf{W}_Q\mathbf{S})$, the second term can be deterministically bounded as

$$\begin{aligned}
&\mathbb{E}_n\left|f_Q(\mathbf{x};\mathbf{W}_L + \mathbf{W}_Q\mathbf{S}) - f_Q(\mathbf{x};\mathbf{W}_Q\mathbf{S})\right|\\
\le\ &\frac{1}{2\sqrt{m}}\sum_{r=1}^m \mathbb{E}_n\left|\sigma''(\mathbf{w}_{0,r}^T\mathbf{x})\left((\sigma_r(\mathbf{w}_Q)_r^T\mathbf{x} + (\mathbf{w}_L)_r^T\mathbf{x})^2 - ((\mathbf{w}_Q)_r^T\mathbf{x})^2\right)\right|\\
\le\ &\frac{1}{2\sqrt{m}}\sum_{r=1}^m \mathbb{E}_n\left|(\mathbf{w}_L)_r^T\mathbf{x}\mathbf{x}^T(2\sigma_r(\mathbf{w}_Q)_r + (\mathbf{w}_L)_r)\right|\\
\le\ &\frac{d}{\sqrt{m}}\sum_{r=1}^m \|(\mathbf{w}_L)_r\|_2(\|(\mathbf{w}_L)_r\|_2 + \|(\mathbf{w}_Q)_r\|_2)\\
\le\ &\frac{d}{\sqrt{m}}(\|\mathbf{W}_L\|_F^2 + \|\mathbf{W}_L\|_F\|\mathbf{W}_Q\|_F)\\
\lesssim\ &m^{-\frac{1}{4}}d^{\frac{3k+1}{4}}.
\end{aligned}$$

We next bound the first term in expectation using Lemma 19:

$$\begin{aligned}
\mathbb{E}_\mathbf{S}\mathbb{E}_n\left|f_L(\mathbf{x};\mathbf{W}_Q\mathbf{S})\right| &\le \left(\mathbb{E}_\mathbf{S}\mathbb{E}_n\left(f_L(\mathbf{x};\mathbf{W}_Q\mathbf{S})\right)^2\right)^{1/2}\\
&\le \frac{1}{\sqrt{m}}\|\mathbf{W}_Q\|_F\\
&\lesssim m^{-\frac{1}{4}}d^{\frac{k-1}{4}}.
\end{aligned}$$

Since the loss is Lipschitz, we can bound the empirical loss as

$$\begin{aligned}
\mathbb{E}_{\mathbf{S}}\hat{L}^Q(\mathbf{W}_S^*) &= \mathbb{E}_{\mathbf{S}}\mathbb{E}_n[\ell(f^*(\mathbf{x}), f_Q(\mathbf{x};\mathbf{W}_{\mathbf{S}}^*) + f_L(\mathbf{x};\mathbf{W}_{\mathbf{S}}^*))] \\
&\leq \mathbb{E}_{\mathbf{S}}\mathbb{E}_n\left|f^*(\mathbf{x}) - f_Q(\mathbf{x};\mathbf{W}_{\mathbf{S}}^*) - f_L(\mathbf{x};\mathbf{W}_{\mathbf{S}}^*)\right| \\
&\leq \mathbb{E}_{\mathbf{S}}\mathbb{E}_n|f_{sp}(\mathbf{x}) - f_Q(\mathbf{x};\mathbf{W}_Q)| \\
&\quad + \mathbb{E}_{\mathbf{S}}\mathbb{E}_n|f_k(\mathbf{x}) - f_L(\mathbf{x};\mathbf{W}_L)| \\
&\quad + \mathbb{E}_{\mathbf{S}}\mathbb{E}_n|f_L(\mathbf{x};\mathbf{W}_{\mathbf{S}}^*) + f_Q(\mathbf{x};\mathbf{W}_{\mathbf{S}}^*) - f_L(\mathbf{x};\mathbf{W}_L) - f_Q(\mathbf{x};\mathbf{W}_Q)| \\
&\lesssim \frac{d^k}{\sqrt{m}} + \sqrt{\frac{d^k}{n}} + m^{-\frac{1}{4}}d^{\frac{3k+1}{4}} \\
&\leq \varepsilon_{min},
\end{aligned}$$

where we used the bounds in Lemmas 1, 2, along with the lower bounds on m in the assumption of the theorem.

Bounding the Regularizers. We begin with $\mathcal{R}_1$:

$$\begin{aligned}
\mathcal{R}_1(\mathbf{W}_S^*) &= \|f_L(\mathbf{x};\mathbf{P}_{>k}\mathbf{W}_L + \mathbf{P}_{>k}\mathbf{W}_Q\mathbf{S})\|_{L^2}^2 \\
&\leq \|f_L(\mathbf{x};\mathbf{P}_{>k}\mathbf{W}_Q\mathbf{S})\|_{L^2}^2 \\
&\leq \|f_L(\mathbf{x};\mathbf{W}_Q\mathbf{S})\|_{L^2}^2,
\end{aligned}$$

since our construction guarantees that $\mathbf{P}_{>k}\mathbf{W}_L = \mathbf{0}$.

By Lemma 19,

$$\mathbb{E}_{\mathbf{S}}\|f_L(\mathbf{x};\mathbf{W}_Q\mathbf{S})\|_{L^2}^2 \leq \frac{1}{m}\|\mathbf{W}_Q\|_F^2 \lesssim m^{-\frac{1}{2}}d^{\frac{k-1}{2}}.$$

Therefore

$$\mathbb{E}_{\mathbf{S}}\mathcal{R}_1(\mathbf{W}_{\mathbf{S}}^*) \lesssim m^{-\frac{1}{2}}d^{\frac{k-1}{2}}.$$

$\mathcal{R}_2$ can be bounded as

$$\begin{aligned}
\mathcal{R}_2(\mathbf{W}_{\mathbf{S}}^*) &= \|f_L(\cdot;\mathbf{P}_{\leq k}\mathbf{W}_{\mathbf{S}}^*)\|_{L^2}^2 \\
&\leq \|f_L(\cdot;\mathbf{W}_{\mathbf{S}}^*)\|_{L^2}^2 \\
&\lesssim \|f_L(\cdot;\mathbf{W}_L)\|_{L^2}^2 + \|f_L(\cdot;\mathbf{W}_Q\mathbf{S})\|_{L^2}^2
\end{aligned}$$

By Lemma 20,

$$\begin{aligned}
\|f_L(\cdot;\mathbf{W}_L)\|_{L^2}^2 &\lesssim \mathbb{E}_n\left[(f_L(\mathbf{x};\mathbf{W}_L))^2\right] \\
&\lesssim \mathbb{E}_n\left[f_k(\mathbf{x})^2\right] + \mathbb{E}_n\left[(f_L(\mathbf{x};\mathbf{W}_L) - f_k(\mathbf{x}))^2\right] \\
&\lesssim 1 + \frac{d^k}{n} \lesssim 1,
\end{aligned}$$

where the last step uses Lemma 16. Therefore, again applying Lemma 19,

$$\mathbb{E}_{\mathbf{S}}\mathcal{R}_2(\mathbf{W}_{\mathbf{S}}^*) \lesssim 1 + \mathbb{E}_{\mathbf{S}}\|f_L(\cdot;\mathbf{W}_Q\mathbf{S})\|_{L^2}^2 \lesssim 1 + m^{-\frac{1}{2}}d^{\frac{k-1}{2}} \lesssim 1.$$

$\mathcal{R}_3$ can be bounded as

$$\begin{aligned}
\mathcal{R}_3(\mathbf{W}_{\mathbf{S}}^*) &= \mathbb{E}_n\left[(f_L(x;\mathbf{P}_{>k}\mathbf{W}_L + \mathbf{P}_{>k}\mathbf{W}_Q\mathbf{S})^2\right] \\
&= \mathbb{E}_n\left[(f_L(x;\mathbf{P}_{>k}\mathbf{W}_Q\mathbf{S}))^2\right] \\
&\lesssim \mathbb{E}_n\left[(f_L(x;\mathbf{W}_Q\mathbf{S}))^2\right] + \mathbb{E}_n\left[(f_L(x;P_{\leq k}\mathbf{W}_Q\mathbf{S}))^2\right],
\end{aligned}$$

since our construction guarantees $\mathbf{P}_{>k}\mathbf{W}_L = \mathbf{0}$.

By Lemma 19,

$$\mathbb{E}_{\mathbf{S}}\mathbb{E}_n\left[(f_L(\mathbf{x};\mathbf{W}_Q\mathbf{S}))^2\right] \leq m^{-\frac{1}{2}}d^{\frac{k-1}{2}}.$$

By Lemma 20,

$$\mathbb{E}_n\left[(f_L(x;\mathbf{P}_{\leq k}\mathbf{W}_Q\mathbf{S}))^2\right] \leq 2\|f_L(\mathbf{x};\mathbf{P}_{\leq k}\mathbf{W}_Q\mathbf{S})\|_{L^2}^2,$$

which can be upper bounded by $2\|f_L(\mathbf{x}; \mathbf{W}_Q\mathbf{S})\|_{L^2}^2$. Therefore

$$\mathbb{E}_{\mathbf{S}}\mathbb{E}_n\left[(f_L(x; \mathbf{P}_{\le k}\mathbf{W}_Q\mathbf{S}))^2\right] \lesssim \mathbb{E}_{\mathbf{S}}\|f_L(x; \mathbf{W}_Q\mathbf{S})\|_{L^2}^2 \lesssim m^{-\frac{1}{2}} d^{\frac{k-1}{2}}.$$

Altogether,

$$\mathbb{E}_{\mathbf{S}}\mathcal{R}_3(\mathbf{W}_{\mathbf{S}}^*) \lesssim m^{-\frac{1}{2}} d^{\frac{k-1}{2}}.$$

Finally, to bound $\mathcal{R}_4$, observe that

$$\|\mathbf{W}_{\mathbf{S}}^*\|_{2,4} \le \|\mathbf{W}_L\|_{2,4} + \|\mathbf{W}_Q\mathbf{S}\|_{2,4} = \|\mathbf{W}_L\|_{2,4} + \|\mathbf{W}_Q\|_{2,4}.$$

By the construction $\|\mathbf{W}_Q\|_{2,4} \le d^{\frac{k-1}{4}}$. Also, since $\mathbf{W}_L \in \text{span}(\mathbf{P}_{\le k})$, by Lemma 21 we have

$$\|\mathbf{W}_L\|_{2,4} \le m^{-\frac{1}{4}} d^{\frac{k}{2}} \|\mathbf{W}_L\|_F \lesssim m^{-\frac{1}{4}} d^{k-\frac{1}{2}}.$$

Therefore $\|\mathbf{W}_{\mathbf{S}}^*\|_{2,4} \lesssim d^{\frac{k-1}{4}}$, and thus

$$\mathcal{R}_4(\mathbf{W}_{\mathbf{S}}^*) \lesssim d^{2(k-1)}.$$

□

Corollary 3. *The solution* $\mathbf{W}^*$ *satisfies*

$$\|\mathbf{W}^*\|_F \lesssim m^{\frac{1}{4}} d^{\frac{k-1}{4}}.$$

Proof. We have

$$\|\mathbf{W}^*\|_F \le \|\mathbf{W}_L^*\|_F + \|\mathbf{W}_Q^*\|_F.$$

By construction $\|\mathbf{W}_L^*\|_F \lesssim d^{\frac{k-1}{2}}$, and also

$$\|\mathbf{W}_Q^*\|_F \le m^{\frac{1}{4}} \|\mathbf{W}_Q^*\|_{2,4} \lesssim m^{\frac{1}{4}} d^{\frac{k-1}{4}}.$$

The desired claim follows since we assume $m \gg d^{k-1}$. □

Corollary 4.

$$\|\mathbf{W}^*\|_{2,\infty} \lesssim m^{-1/4} d^{\frac{k-1}{2}}$$

Proof. We can writte

$$\|\mathbf{W}^*\|_{2,\infty} \le \|\mathbf{W}_L^*\|_{2,\infty} + \|\mathbf{W}_Q^*\mathbf{S}^*\|_{2,\infty} = \|\mathbf{W}_L^*\|_{2,\infty} + \|\mathbf{W}_Q^*\|_{2,\infty}.$$

By Corollary 2, $\|\mathbf{W}_Q^*\|_{2,\infty} \lesssim m^{-1/4} d^{\frac{k-1}{2}}$. Also, since $\mathbf{W}_L^* \in \text{span}(\mathbf{P}_{\le k})$, we can apply Lemma 21 to get $\|\mathbf{W}_L^*\|_{2,\infty} \le m^{-\frac{1}{2}} d^{\frac{k}{2}} \|\mathbf{W} *_L\|_F \le m^{-\frac{1}{2}} d^{k-\frac{1}{2}}$. Altogether, since $m \ge d^{2k}$, we get

$$\|\mathbf{W}^*\|_{2,\infty} \lesssim m^{-1/4} d^{\frac{k-1}{2}}.$$

□

B.4 Expressivity Helper Lemmas

Lemma 16. *With probability* $1 - d^{-10}$,

$$\mathbb{E}_n[f_k(\mathbf{x})^2] \lesssim 1.$$

Proof. Since $\|f_k\|_{L^2} = 1$ and f is degree k, $|f_k(\mathbf{x})|^2 \le n_k$ for all $\mathbf{x}$. Furthermore,

$$\mathbb{E}_\mu[(|f_k(\mathbf{x})|^2 - 1)^2] \le \mathbb{E}_\mu[f_k(\mathbf{x})^4] \le n_k.$$

Therefore by Bernstein's Inequality, with probability $1 - d^{-10}$ we have

$$\mathbb{E}_n[f_k(\mathbf{x})^2] - 1 \le C\left(\sqrt{\frac{n_k \log d}{n}} + \frac{n_k \log d}{n}\right) \lesssim 1,$$

since $n \gtrsim d^k \log d \gtrsim n_k \log d$ □

Lemma 17 ([42, Exercise 4.7.3]). *With probability* $1 - d^{-10}$,

$$\left\|\mathbb{E}_n \mathbf{x}\mathbf{x}^T - \mathbf{I}_d\right\|_{op} \leq \frac{1}{2}$$

Lemma 18. *Recall*

$$\mathbf{\Sigma}_{\leq n_k} := \mathbb{E}_\mu \left[\varphi(\mathbf{x})^T \mathbf{P}_{\leq k} \varphi(\mathbf{x})\right] = \sum_{i=1}^{n_k} \lambda_i \mathbf{v}_i \mathbf{v}_i^T.$$

With probability $1 - d^{-10}$,

$$\left\|\mathbb{E}_n \left[(\mathbf{\Sigma}_{\leq n_k}^\dagger)^{\frac{1}{2}} \varphi(\mathbf{x})\varphi(\mathbf{x})^T (\mathbf{\Sigma}_{\leq n_k}^\dagger)^{\frac{1}{2}}\right] - \mathbf{P}_{\leq k}\right\|_{op} \leq \frac{1}{2}$$

Proof. Observe that

$$\begin{aligned}
\mathbb{E}_\mu \left[(\mathbf{\Sigma}_{\leq n_k}^\dagger)^{\frac{1}{2}} \varphi(\mathbf{x})\varphi(\mathbf{x})^T (\mathbf{\Sigma}_{\leq n_k}^\dagger)^{\frac{1}{2}}\right] &= (\mathbf{\Sigma}_{\leq n_k}^\dagger)^{\frac{1}{2}} \mathbb{E}_\mu \left[\varphi(\mathbf{x})\varphi(\mathbf{x})^T\right] (\mathbf{\Sigma}_{\leq n_k}^\dagger)^{\frac{1}{2}} \\
&= (\mathbf{\Sigma}_{\leq n_k}^\dagger)^{\frac{1}{2}} \mathbf{\Sigma} (\mathbf{\Sigma}_{\leq n_k}^\dagger)^{\frac{1}{2}} \\
&= \mathbf{P}_{\leq k}.
\end{aligned}$$

Therefore by [42, 5.6.4], with probability $1 - d^{-10}$

$$\begin{aligned}
\left\|\mathbb{E}_n \left[(\mathbf{\Sigma}_{\leq n_k}^\dagger)^{\frac{1}{2}} \varphi(\mathbf{x})\varphi(\mathbf{x})^T (\mathbf{\Sigma}_{\leq n_k}^\dagger)^{\frac{1}{2}}\right] - \mathbf{P}_{\leq k}\right\|_{op} &\leq C\left(\sqrt{\frac{d^k \log d}{n}} + \frac{d^k \log d}{n}\right) \\
&\leq \frac{1}{2},
\end{aligned}$$

since $n \gtrsim d^k \log d$. □

Lemma 19. *On the event where Lemma 17 holds, for all* $\mathbf{W}$ *we have*

$$\mathbb{E}_{\mathbf{S}} \mathbb{E}_n [(f_L(\mathbf{x}; \mathbf{WS}))^2] \lesssim \frac{1}{m} \|\mathbf{W}\|_F^2.$$

Furthermore, we have

$$\mathbb{E}_{\mathbf{S}} \|f_L(\mathbf{x}; \mathbf{WS})\|_{L^2}^2 = \mathbb{E}_{\mathbf{S}} \mathbb{E}_\mu [(f_L(\mathbf{x}; \mathbf{WS}))^2] \lesssim \frac{1}{m} \|\mathbf{W}\|_F^2.$$

Proof. We have

$$\begin{aligned}
\mathbb{E}_{\mathbf{S}} \mathbb{E}_n [(f_L(\mathbf{x}; \mathbf{WS}))^2] &= \mathbb{E}_{\mathbf{S}} \mathbb{E}_n \left(\frac{1}{\sqrt{m}} \sum_{r=1}^m \sigma'(\mathbf{w}_{0,r}^T \mathbf{x}) \mathbf{x}^T \mathbf{w}_r \sigma_r\right)^2 \\
&\leq \mathbb{E}_n \frac{1}{m} \sum_{r=1}^m (\mathbf{w}_r^T \mathbf{x})^2 \\
&\leq \frac{1}{m} \sum_{r=1}^m \mathbf{w}_r^T \mathbb{E}_n [\mathbf{x}\mathbf{x}^T] \mathbf{w}_r \\
&\lesssim \frac{1}{m} \|\mathbf{W}\|_F^2
\end{aligned}$$

The proof in the population case is identical. □

Lemma 20. *On the event where Lemma 18 holds,*

$$\frac{1}{2} \|f_L(\mathbf{x}; \mathbf{P}_{\leq k} \mathbf{W})\|_{L^2}^2 \leq \mathbb{E}_n [(f_L(\mathbf{x}; \mathbf{P}_{\leq k} \mathbf{W}))^2] \leq \frac{3}{2} \|f_L(\mathbf{x}; \mathbf{P}_{\leq k} \mathbf{W})\|_{L^2}^2.$$

for all $\mathbf{W}$.

Proof. We can write

$$\begin{aligned}
&\mathbb{E}_n\left[(f_L(x;\mathbf{P}_{\le k}\mathbf{W}))^2\right]\\
=\;&\mathrm{vec}(\mathbf{W})^T\mathbb{E}_n\left[\mathbf{P}_{\le k}\varphi(\mathbf{x})\varphi(\mathbf{x})^T\mathbf{P}_{\le k}\right]\mathrm{vec}(\mathbf{W})\\
=\;&\mathrm{vec}(\mathbf{W})^T\mathbf{\Sigma}_{\le n_k}^{\frac{1}{2}}\mathbb{E}_n\left[(\mathbf{\Sigma}_{\le n_k}^{\dagger})^{\frac{1}{2}}\varphi(\mathbf{x})\varphi(\mathbf{x})^T(\mathbf{\Sigma}_{\le n_k}^{\dagger})^{\frac{1}{2}}\right]\mathbf{\Sigma}_{\le n_k}^{\frac{1}{2}}\mathrm{vec}(\mathbf{W})\\
=\;&\mathrm{vec}(\mathbf{W})^T\mathbf{\Sigma}_{\le n_k}^{\frac{1}{2}}\left(\mathbf{P}_{\le k}+\mathbb{E}_n\left[(\mathbf{\Sigma}_{\le n_k}^{\dagger})^{\frac{1}{2}}\varphi(\mathbf{x})\varphi(\mathbf{x})^T(\mathbf{\Sigma}_{\le n_k}^{\dagger})^{\frac{1}{2}}\right]-\mathbf{P}_{\le k}\right)\mathbf{\Sigma}_{\le n_k}^{\frac{1}{2}}\mathrm{vec}(\mathbf{W}).
\end{aligned}$$

Therefore, by Lemma 18,

$$\begin{aligned}
&\left|\mathbb{E}_n[(f_L(\mathbf{x};\mathbf{P}_{\le k}\mathbf{W}))^2]-\|f_L(\mathbf{x};\mathbf{P}_{\le k}\mathbf{W})\|_{L^2}^2\right|\\
&=\left|\mathrm{vec}(\mathbf{W})^T\mathbf{\Sigma}_{\le n_k}^{\frac{1}{2}}\left(\mathbb{E}_n\left[(\mathbf{\Sigma}_{\le n_k}^{\dagger})^{\frac{1}{2}}\varphi(\mathbf{x})\varphi(\mathbf{x})^T(\mathbf{\Sigma}_{\le n_k}^{\dagger})^{\frac{1}{2}}\right]-\mathbf{P}_{\le k}\right)\mathbf{\Sigma}_{\le n_k}^{\frac{1}{2}}\mathrm{vec}(\mathbf{W})\right|\\
&\le\mathrm{vec}(\mathbf{W})^T\mathbf{\Sigma}_{\le n_k}\mathrm{vec}(\mathbf{W})\left(\left\|\mathbb{E}_n\left[(\mathbf{\Sigma}_{\le n_k}^{\dagger})^{\frac{1}{2}}\varphi(\mathbf{x})\varphi(\mathbf{x})^T(\mathbf{\Sigma}_{\le n_k}^{\dagger})^{\frac{1}{2}}\right]-\mathbf{P}_{\le k}\right\|_{op}\right)\\
&\le\frac{1}{2}\mathrm{vec}(\mathbf{W})^T\mathbf{\Sigma}_{\le n_k}\mathrm{vec}(\mathbf{W})\\
&=\frac{1}{2}\|f_L(\mathbf{x};\mathbf{P}_{\le k}\mathbf{W})\|_{L^2}^2,
\end{aligned}$$

as desired. □

Lemma 21. *For any* $\mathbf{W}\in$ *span*$(\mathbf{P}_{\le k})$*, we have*

$$\|\mathbf{W}\|_{2,\infty}\le m^{-\frac{1}{2}}d^{\frac{k}{2}}\|\mathbf{W}\|_F$$

and

$$\|\mathbf{W}\|_{2,4}\le m^{-\frac{1}{4}}d^{\frac{k}{2}}\|\mathbf{W}\|_F$$

Proof. For $(r,s)\in[m]\times[d]$, let $\mathbf{e}_{(r,s)}$ denote the $((d-1)r+s)$th canonical basis vector in $\mathbb{R}^{md}$, so that $\mathbf{e}_{(r,s)}^T\mathrm{vec}(\mathbf{W})=\{\mathbf{w}_r\}_s$. Let $c_1,\dots,c_{n_k}$ be scalars such that $\sum_{i=1}^{n_k}c_i^2=\|\mathbf{W}\|_F^2$ and

$$\mathrm{vec}(\mathbf{W})=\sum_{i=1}^{n_k}c_i\mathbf{v}_i.$$

By Cauchy, we can bound

$$\left|\langle\mathbf{W},\mathbf{e}_{(r,s)}\rangle\right|\le\|\mathbf{W}\|_{\mathbf{\Sigma}^{-1}}\|e_{(r,s)}\|_{\mathbf{\Sigma}}.$$

Observe that

$$\|\mathbf{W}\|_{\mathbf{\Sigma}^{-1}}^2=\sum_{i=1}^{n_k}c_i^2\lambda_i^{-1}\le\lambda_{n_k}^{-1}\|\mathbf{W}\|_F^2.$$

Furthermore, since $\mathbf{e}_{(r,s)}^T\varphi(\mathbf{x})=\frac{1}{\sqrt{m}}\sigma'(\mathbf{w}_{0,r}^T\mathbf{x})\mathbf{x}_s$

$$\|\mathbf{e}_{(r,s)}\|_{\Sigma}^2=\mathbb{E}_\mu\left[(\mathbf{e}_{(r,s)}^T\varphi(\mathbf{x}))^2\right]\le\frac{1}{m}\mathbb{E}_\mu[\mathbf{x}_s^2]\le\frac{1}{m}.$$

By Lemma 15 $\lambda_{n_k}=\Theta(d^{-k+1})$, and thus we can bound

$$|\langle\mathbf{W},\mathbf{e}(r,s)\rangle|\le m^{-\frac{1}{2}}\lambda_{n_k}^{-\frac{1}{2}}\|\mathbf{W}\|_F\le m^{-\frac{1}{2}}d^{\frac{k-1}{2}}\|\mathbf{W}\|_F.$$

Thus every row $\mathbf{w}_r$ satisfies

$$\|\mathbf{w}_r\|_2\le m^{-\frac{1}{2}}d^{\frac{k}{2}}\|\mathbf{W}\|_F,$$

so

$$\|\mathbf{W}\|_{2,4}\le m^{\frac{1}{4}}\cdot m^{-\frac{1}{2}}d^{\frac{k}{2}}\|\mathbf{W}\|_F\le m^{-\frac{1}{4}}d^{\frac{k}{2}}\|\mathbf{W}\|_F.$$

□

C Landscape Proofs

C.1 Coupling Lemmas

Recall that $\hat{L}^Q(\mathbf{W}) = \mathbb{E}_n[\ell(y, f_L(\mathbf{x}; \mathbf{W}) + f_Q(\mathbf{x}; \mathbf{W}))]$ denotes the empirical loss of the quadratic model. As in [7], we begin by showing the losses, gradients, and Hessians for $\hat{L}^Q(\mathbf{W})$ and $\hat{L}(\mathbf{W})$ are close for $\mathbf{W}$ satisfying a norm bound. This is given by the following 3 coupling lemmas.

Lemma 22 (Coupling of Losses)**.**

$$\left|\hat{L}(\mathbf{W}) - \hat{L}^Q(\mathbf{W})\right| \lesssim d^{3/2} m^{-1/4} \|\mathbf{W}\|_{2,4}^3$$

Lemma 23 (Coupling of Gradients)**.**

$$\left|\langle \nabla \hat{L}(\mathbf{W}), \tilde{\mathbf{W}} \rangle - \langle \nabla \hat{L}^Q(\mathbf{W}), \tilde{\mathbf{W}} \rangle\right| \lesssim d^{3/2} m^{-1/4} (\|\mathbf{W}\|_{2,4}^3 + \|\tilde{\mathbf{W}}\|_{2,4}^3) \cdot \max_{i \in [n]} \left|\langle \nabla_{\mathbf{W}} f(\mathbf{x}; \mathbf{W}), \tilde{\mathbf{W}} \rangle\right|$$

Lemma 24 (Coupling of Hessians)**.**

$$\begin{aligned}
&\left|\nabla^2 \hat{L}(\mathbf{W})[\tilde{\mathbf{W}}, \tilde{\mathbf{W}}] - \nabla^2 \hat{L}^Q(\mathbf{W})[\tilde{\mathbf{W}}, \tilde{\mathbf{W}}]\right| \\
&\lesssim d^{3/2} m^{-\frac{1}{4}} \left(\|\mathbf{W}\|_{2,4}^3 + \|\tilde{\mathbf{W}}\|_{2,4}^3\right) \left(d \|\tilde{\mathbf{W}}\|_{2,4}^2 + \max_{i \in [n]} \left|\langle \nabla_{\mathbf{W}} f(\mathbf{x}_i; \mathbf{W}), \tilde{\mathbf{W}} \rangle\right|^2 + \max_{i \in [n]} \left|\langle \nabla f(\mathbf{x}_i, \mathbf{W}), \tilde{\mathbf{W}} \rangle\right|\right) \\
&\quad + d^3 \|\mathbf{W}\|_{2,4}^4 \|\tilde{\mathbf{W}}\|_{2,\infty}^2.
\end{aligned}$$

While these Lemmas are similar in structure to those in [7], extra care must be taken to properly deal with the effect of the f_L terms. These proofs are presented in Appendix C.1.2.

C.1.1 Auxiliary Results

We first prove some intermediate results which are used in the proofs of the coupling lemmas:

Lemma 25 (Coupling of Function Values)**.**

$$|f(\mathbf{x}; \mathbf{W}) - f_L(\mathbf{x}; \mathbf{W}) - f_Q(\mathbf{x}; \mathbf{W})| \leq \frac{1}{\sqrt{m}} \sum_{r=1}^m |\mathbf{w}_r^T \mathbf{x}|^3$$

Proof.

$$\begin{aligned}
&|f(\mathbf{x}; \mathbf{W}) - f_L(\mathbf{x}; \mathbf{W}) - f_Q(\mathbf{x}; \mathbf{W})| \\
&= \frac{1}{\sqrt{m}} \left|\sum_{r=1}^m \sigma(\mathbf{w}_{0,r}^T \mathbf{x} + \mathbf{w}_r^T \mathbf{x}) - \sigma(\mathbf{w}_{0,r}^T \mathbf{x}) - \sigma'(\mathbf{w}_{0,r}^T \mathbf{x})(\mathbf{w}_r^T \mathbf{x}) - \frac{1}{2}\sigma''(\mathbf{w}_{0,r}^T \mathbf{x})(\mathbf{w}_r^T \mathbf{x})^2\right| \\
&\leq \frac{1}{\sqrt{m}} \sum_{r=1}^m |\sigma'''|_\infty |\mathbf{w}_r^T \mathbf{x}|^3 \\
&\leq \frac{1}{\sqrt{m}} \sum_{r=1}^m |\mathbf{w}_r^T \mathbf{x}|^3.
\end{aligned}$$

□

Lemma 26 (Coupling of Function Gradients)**.**

$$\left|\langle \nabla_{\mathbf{W}} f_L(\mathbf{x}; \mathbf{W}) + \nabla_{\mathbf{W}} f_Q(\mathbf{x}; \mathbf{W}), \tilde{\mathbf{W}} \rangle - \langle \nabla_{\mathbf{W}} f(\mathbf{x}; \mathbf{W}), \tilde{\mathbf{W}} \rangle\right| \leq \frac{1}{\sqrt{m}} \sum_{r=1}^m |\tilde{\mathbf{w}}_r^T \mathbf{x}| |\mathbf{w}_r^T \mathbf{x}|^2$$

Proof. Taylor expanding σ', we have

$$\left|\sigma'(\mathbf{w}_{0,r}^T\mathbf{x}+\mathbf{w}_r^T\mathbf{x})-\sigma'(\mathbf{w}_{0,r}^T\mathbf{x})+\sigma''(\mathbf{w}_{0,r}^T\mathbf{x})(\mathbf{w}_r^T\mathbf{x})\right|\leq|\mathbf{w}_r^T\mathbf{x}|^2.$$

Therefore

$$\begin{aligned}
&\left|\langle\nabla_{\mathbf{W}}f_L(\mathbf{x};\mathbf{W})+\nabla_{\mathbf{W}}f_Q(\mathbf{x};\mathbf{W}),\tilde{\mathbf{W}}\rangle-\langle\nabla_{\mathbf{W}}f(\mathbf{x};\mathbf{W}),\tilde{\mathbf{W}}\rangle\right|\\
=&\left|\frac{1}{\sqrt{m}}\sum_{r=1}^m a_r\left(\sigma'(\mathbf{w}_{0,r}^T\mathbf{x})(\tilde{\mathbf{w}}_r^T\mathbf{x})+\sigma''(\mathbf{w}_{0,r}^T\mathbf{x})(\mathbf{w}_r^T\mathbf{x})(\tilde{\mathbf{w}}_r^T\mathbf{x})-\sigma'(\mathbf{w}_{0,r}^T\mathbf{x}+\mathbf{w}_r^T\mathbf{x})(\tilde{\mathbf{w}}_r^T\mathbf{x})\right)\right|\\
\leq&\frac{1}{\sqrt{m}}\sum_{r=1}^m|\tilde{\mathbf{w}}_r^T\mathbf{x}|\left|\sigma'(\mathbf{w}_{0,r}^T\mathbf{x}+\mathbf{w}_r^T\mathbf{x})-\sigma'(\mathbf{w}_{0,r}^T\mathbf{x})+\sigma''(\mathbf{w}_{0,r}^T\mathbf{x})(\mathbf{w}_r^T\mathbf{x})\right|\\
\leq&\frac{1}{\sqrt{m}}\sum_{r=1}^m|\tilde{\mathbf{w}}_r^T\mathbf{x}||\mathbf{w}_r^T\mathbf{x}|^2.
\end{aligned}$$

□

Lemma 27.

$$\max_{i\in[n]}\left|\langle\nabla_{\mathbf{W}}f(\mathbf{x}_i;\mathbf{W}),\tilde{\mathbf{W}}\rangle\right|\lesssim n^{\frac{1}{2}}\mathbb{E}_n\left[(f_L(\mathbf{x};\tilde{\mathbf{W}}))^2\right]^{\frac{1}{2}}+dn^{\frac{1}{2}}m^{-\frac{1}{2}}\|\mathbf{W}\|_F\|\tilde{\mathbf{W}}\|_F$$

Proof. We can decompose

$$\mathbb{E}_n\left[\langle\nabla_{\mathbf{W}}f(\mathbf{x};\mathbf{W}),\tilde{\mathbf{W}}\rangle^2\right]\leq2\mathbb{E}_n\left[\langle\nabla_{\mathbf{W}}f(\mathbf{x};\mathbf{0}),\tilde{\mathbf{W}}\rangle^2\right]+2\mathbb{E}_n\left[\langle\nabla_{\mathbf{W}}f(\mathbf{x};\mathbf{0})-\nabla_{\mathbf{W}}f(\mathbf{x};\mathbf{W}),\tilde{\mathbf{W}}\rangle^2\right]$$

We can bound

$$\begin{aligned}
\langle\nabla_{\mathbf{W}}f(\mathbf{x};\mathbf{0})-\nabla_{\mathbf{W}}f(\mathbf{x};\mathbf{W}),\tilde{\mathbf{W}}\rangle^2&\leq\|\nabla_{\mathbf{W}}f(\mathbf{x};\mathbf{0})-\nabla_{\mathbf{W}}f(\mathbf{x};\mathbf{W})\|_2^2\|\tilde{\mathbf{W}}\|_F^2\\
&\leq\|\tilde{\mathbf{W}}\|_F^2\sum_{r=1}^m\frac{d}{m}\left|\sigma'(\mathbf{w}_{0,r}^T\mathbf{x})-\sigma'(\mathbf{w}_{0,r}^T\mathbf{x}+\mathbf{w}_r^T\mathbf{x})\right|^2\\
&\leq\|\tilde{\mathbf{W}}\|_F^2\sum_{r=1}^m\frac{d}{m}(\mathbf{w}_r^T\mathbf{x})^2\\
&\leq\frac{d^2}{m}\|\mathbf{W}\|_F^2\|\tilde{\mathbf{W}}\|_F^2.
\end{aligned}$$

Thus

$$\mathbb{E}_n\left[\langle\nabla_{\mathbf{W}}f(\mathbf{x};\mathbf{W}),\tilde{\mathbf{W}}\rangle^2\right]\lesssim\mathbb{E}_n\left[(f_L(\mathbf{x};\tilde{\mathbf{W}}))^2\right]+\frac{d^2}{m}\|\mathbf{W}\|_F^2\|\tilde{\mathbf{W}}\|_F^2,$$

So

$$\begin{aligned}
\max_{i\in[n]}\left|\langle\nabla_{\mathbf{W}}f(\mathbf{x}_i;\mathbf{W}),\tilde{\mathbf{W}}\rangle\right|&\leq\left(\sum_{i=1}^n\langle\nabla_{\mathbf{W}}f(\mathbf{x}_i;\mathbf{W}),\tilde{\mathbf{W}}\rangle^2\right)^{1/2}\\
&\lesssim\left(n\left(\mathbb{E}_n\left[(f_L(\mathbf{x};\tilde{\mathbf{W}}))^2\right]+\frac{d^2}{m}\|\mathbf{W}\|_F^2\|\tilde{\mathbf{W}}\|_F^2\right)\right)^{1/2}\\
&\lesssim n^{\frac{1}{2}}\mathbb{E}_n\left[(f_L(\mathbf{x};\tilde{\mathbf{W}}))^2\right]^{\frac{1}{2}}+dn^{\frac{1}{2}}m^{-\frac{1}{2}}\|\mathbf{W}\|_F\|\tilde{\mathbf{W}}\|_F.
\end{aligned}$$

□

C.1.2 Proof of Coupling Lemmas

Proof of Lemma 22. By Lipschitzness of the loss,

$$\begin{aligned}|\hat{L}(\mathbf{W}) - \hat{L}^Q(\mathbf{W})| &\le \mathbb{E}_n\left|\ell(y, f(\mathbf{x};\mathbf{W})) - \ell(y, f_L(\mathbf{x};\mathbf{W}) + f_Q(\mathbf{x};\mathbf{W}))\right| \\ &\le \mathbb{E}_n\left|f(\mathbf{x};\mathbf{W}) - f_L(\mathbf{x};\mathbf{W}) - f_Q(\mathbf{x};\mathbf{W})\right| \\ &\le \frac{1}{\sqrt{m}}\sum_{r=1}^m \mathbb{E}_n|\mathbf{w}_r^T\mathbf{x}|^3 \\ &\le C\frac{d^{3/2}}{\sqrt{m}}\sum_{r=1}^m \|\mathbf{w}_r\|_2^3 \\ &\le Cm^{-1/4}d^{3/2}\|\mathbf{W}\|_{2,4}^3.\end{aligned}$$

□

Proof of Lemma 23. The gradients are

$$\begin{aligned}\langle\nabla\hat{L}(\mathbf{W}), \tilde{\mathbf{W}}\rangle &= \mathbb{E}_n\left[\ell'(y, f(\mathbf{x};\mathbf{W}))\cdot\langle\nabla_{\mathbf{W}} f(\mathbf{x};\mathbf{W}), \tilde{\mathbf{W}}\rangle\right] \\ \langle\nabla\hat{L}^Q(\mathbf{W}), \tilde{\mathbf{W}}\rangle &= \mathbb{E}_n\left[\ell'(y, f_L(\mathbf{x};\mathbf{W}) + f_Q(\mathbf{x};\mathbf{W}))\cdot\langle\nabla_{\mathbf{W}} f_L(\mathbf{x};\mathbf{W}) + \nabla_{\mathbf{W}} f_Q(\mathbf{x};\mathbf{W}), \tilde{\mathbf{W}}\rangle\right]\end{aligned}$$

Therefore

$$\begin{aligned}&\left|\langle\nabla\hat{L}(\mathbf{W}), \tilde{\mathbf{W}}\rangle - \langle\nabla\hat{L}^Q(\mathbf{W}), \tilde{\mathbf{W}}\rangle\right| \\ &\le \mathbb{E}_n\left[|\ell'(y, f_L(\mathbf{x};\mathbf{W}) + f_Q(\mathbf{x};\mathbf{W}))|\cdot\left|\langle\nabla_{\mathbf{W}} f_L(\mathbf{x};\mathbf{W}) + \nabla_{\mathbf{W}} f_Q(\mathbf{x};\mathbf{W}), \tilde{\mathbf{W}}\rangle - \langle\nabla_{\mathbf{W}} f(\mathbf{x};\mathbf{W}), \tilde{\mathbf{W}}\rangle\right|\right] \\ &+ \mathbb{E}_n\left[|\ell'(y, f_L(\mathbf{x};\mathbf{W}) + f_Q(\mathbf{x};\mathbf{W})) - \ell'(y, f(\mathbf{x};\mathbf{W}))|\cdot\left|\langle\nabla_{\mathbf{W}} f(\mathbf{x};\mathbf{W}), \tilde{\mathbf{W}}\rangle\right|\right] \\ &\le \frac{1}{\sqrt{m}}\sum_{r=1}^m \mathbb{E}_n\left[|\tilde{\mathbf{w}}_r^T\mathbf{x}||\mathbf{w}_r^T\mathbf{x}|^2\right] + \mathbb{E}_n\left[|f(\mathbf{x};\mathbf{W}) - f_L(\mathbf{x};\mathbf{W}) - f_Q(\mathbf{x};\mathbf{W})|\cdot\left|\langle\nabla_{\mathbf{W}} f(\mathbf{x};\mathbf{W}), \tilde{\mathbf{W}}\rangle\right|\right],\end{aligned}$$

by Lemma 26. The first term is

$$\begin{aligned}\frac{1}{\sqrt{m}}\sum_{r=1}^m \mathbb{E}_n\left[|\tilde{\mathbf{w}}_r^T\mathbf{x}||\mathbf{w}_r^T\mathbf{x}|^2\right] &\le \frac{1}{\sqrt{m}}\sum_{r=1}^m \mathbb{E}_n\left[\frac{1}{3}|\tilde{\mathbf{w}}_r^T\mathbf{x}|^3 + \frac{2}{3}|\mathbf{w}_r^T\mathbf{x}|^3\right] \\ &\le \frac{Cd^{3/2}}{\sqrt{m}}\sum_{r=1}^m \|\mathbf{w}_r\|_2^3 + \|\tilde{\mathbf{w}}_r\|_2^3 \\ &\le Cd^{3/2}m^{-1/4}\left(\|\mathbf{W}\|_{2,4}^3 + \|\tilde{\mathbf{W}}\|_{2,4}^3\right).\end{aligned}$$

The second term can be bounded as

$$\begin{aligned}&\mathbb{E}_n\left[|f(\mathbf{x};\mathbf{W}) - f_L(\mathbf{x};\mathbf{W}) - f_Q(\mathbf{x};\mathbf{W})|\cdot\langle\nabla_{\mathbf{W}} f(\mathbf{x};\mathbf{W}), \tilde{\mathbf{W}}\rangle\right] \\ \le\ &\mathbb{E}_n\left|f(\mathbf{x};\mathbf{W}) - f_L(\mathbf{x};\mathbf{W}) - f_Q(\mathbf{x};\mathbf{W})\right|\cdot\max_{i\in[n]}\left|\langle\nabla_{\mathbf{W}} f(\mathbf{x}_i;\mathbf{W}), \tilde{\mathbf{W}}\rangle\right| \\ \le\ &Cd^{3/2}m^{-1/4}\|\mathbf{W}\|_{2,4}^3\cdot\max_{i\in[n]}\left|\langle\nabla_{\mathbf{W}} f(\mathbf{x};\mathbf{W}), \tilde{\mathbf{W}}\rangle\right|,\end{aligned}$$

by Lemma 25.

□

Proof of Lemma 24. The Hessians are

$$\begin{aligned}&\nabla^2\hat{L}^Q(\mathbf{W})[\tilde{\mathbf{W}}, \tilde{\mathbf{W}}] \\ =\ &\mathbb{E}_n\left[\ell'(y, f_L(\mathbf{x};\mathbf{W}) + f_Q(\mathbf{x};\mathbf{W}))\frac{1}{\sqrt{m}}\sum_{r=1}^m a_r\sigma''(\mathbf{w}_{0,r}^T\mathbf{x})(\tilde{\mathbf{w}}_r^T\mathbf{x})^2\right] \\ &+ \mathbb{E}_n\left[\ell''(y, f_L(\mathbf{x};\mathbf{W}) + f_Q(\mathbf{x};\mathbf{W}))\left(\langle\nabla_{\mathbf{W}} f_L(\mathbf{x};\mathbf{W}) + \nabla_{\mathbf{W}} f_Q(\mathbf{x};\mathbf{W}), \tilde{\mathbf{W}}\rangle\right)^2\right]\end{aligned}$$

and

$$\begin{aligned}\nabla^2 \hat{L}(\mathbf{W})[\tilde{\mathbf{W}}, \tilde{\mathbf{W}}] = \ & \mathbb{E}_n \left[\ell'(y, f(\mathbf{x}; \mathbf{W})) \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r \sigma''((\mathbf{w}_{0,r} + \mathbf{w}_r)^T x)(\tilde{\mathbf{w}}_r^T \mathbf{x})^2 \right] \\ & + \mathbb{E}_n \left[\ell''(y, f(\mathbf{x}; \mathbf{W})) \langle \nabla_{\mathbf{W}} f(\mathbf{x}; \mathbf{W}), \tilde{\mathbf{W}} \rangle^2 \right]\end{aligned}$$

The difference between the first terms can be bounded by

$$\begin{aligned}& \mathbb{E}_n \left[(\ell'(y, f_L(\mathbf{x}; \mathbf{W}) + f_Q(\mathbf{x}; \mathbf{W})) - \ell'(y, f(\mathbf{x}; \mathbf{W}))) \sum_{r=1}^{m} a_r \sigma''(\mathbf{w}_{0,r}^T \mathbf{x})(\tilde{\mathbf{w}}_r^T \mathbf{x})^2 \right] \\ & + \mathbb{E}_n \left[\frac{1}{\sqrt{m}} \sum_{r=1}^{m} \left| (\tilde{\mathbf{w}}_r^T \mathbf{x})^2 \left(\sigma''(\mathbf{w}_{0,r}^T \mathbf{x}) - \sigma''(\mathbf{w}_{0,r}^T \mathbf{x} + \mathbf{w}_r^T \mathbf{x}) \right) \right| \right] \\ \leq \ & \mathbb{E}_n \left[|f_L(\mathbf{x}; \mathbf{W}) + f_Q(\mathbf{x}; \mathbf{W}) - f(\mathbf{x}; \mathbf{W})| \frac{1}{\sqrt{m}} \sum_{r=1}^{m} |\tilde{\mathbf{w}}_r^T \mathbf{x}|^2 \right] + \frac{1}{\sqrt{m}} \mathbb{E}_n \left[\sum_{r=1}^{m} |\mathbf{w}_r^T \mathbf{x}| |\tilde{\mathbf{w}}_r^T \mathbf{x}|^2 \right] \\ \leq \ & \frac{1}{m} \mathbb{E}_n \left[\left(\sum_{r=1}^{m} |\mathbf{w}_r^T \mathbf{x}|^3 \right) \left(\sum_{r=1}^{m} |\tilde{\mathbf{w}}_r^T \mathbf{x}|^2 \right) \right] + \frac{1}{\sqrt{m}} \mathbb{E}_n \left[\sum_{r=1}^{m} |\mathbf{w}_r^T x| |\tilde{\mathbf{w}}_r^T x|^2 \right] \\ \leq \ & \frac{1}{m} d^{3/2} \left(\sum_{r=1}^{m} \|\mathbf{w}_r\|_2^3 \right) \mathbb{E}_n \left[\sum_{r=1}^{m} |\tilde{\mathbf{w}}_r^T \mathbf{x}|^2 \right] + \frac{1}{\sqrt{m}} \mathbb{E}_n \left[\sum_{r=1}^{m} \frac{1}{3} |\mathbf{w}_r^T x|^3 + \frac{2}{3} |\tilde{\mathbf{w}}_r^T x|^3 \right] \\ \leq \ & \frac{d^{5/2}}{m} \left(\sum_{r=1}^{m} \|\mathbf{w}_r\|_2^3 \right) \left(\sum_{r=1}^{m} \|\tilde{\mathbf{w}}_r\|_2^2 \right) + \frac{d^{3/2}}{\sqrt{m}} \sum_{r=1}^{m} \left(\|\mathbf{w}_r\|_2^3 + \|\tilde{\mathbf{w}}_r\|_2^3 \right) \\ \leq \ & \frac{d^{5/2}}{m} \cdot m^{1/4} \|\mathbf{W}\|_{2,4}^3 \cdot m^{1/2} \|\tilde{\mathbf{W}}\|_{2,4}^2 + \frac{d^{3/2}}{\sqrt{m}} m^{1/4} \left(\|\mathbf{W}\|_{2,4}^3 + \|\tilde{\mathbf{W}}\|_{2,4}^3 \right) \\ \leq \ & O\left(d^{5/2} m^{-1/4} \left(\|\mathbf{W}\|_{2,4}^3 + \|\tilde{\mathbf{W}}\|_{2,4}^3 \right) \|\tilde{\mathbf{W}}\|_{2,4}^2 \right).\end{aligned}$$

The difference between the second terms is upper bounded by

$$\begin{aligned}& \mathbb{E}_n \left[(\ell''(y, f_L(\mathbf{x}; \mathbf{W}) + f_Q(\mathbf{x}; \mathbf{W})) - \ell''(y, f(\mathbf{x}; \mathbf{W}))) \langle \nabla_{\mathbf{W}} f(\mathbf{x}; \mathbf{W}), \tilde{\mathbf{W}} \rangle^2 \right] \\ & + \mathbb{E}_n \left| \left(\langle \nabla_{\mathbf{W}} f_L(\mathbf{x}; \mathbf{W}) + \nabla_{\mathbf{W}} f_Q(\mathbf{x}; \mathbf{W}), \tilde{\mathbf{W}} \rangle \right)^2 - \langle \nabla_{\mathbf{W}} f(\mathbf{x}; \mathbf{W}), \tilde{\mathbf{W}} \rangle^2 \right|\end{aligned}$$

The first term can be bounded by

$$\begin{aligned}& \max_{i \in [n]} \langle \nabla_{\mathbf{W}} f(\mathbf{x}_i; \mathbf{W}), \tilde{\mathbf{W}} \rangle^2 \cdot \mathbb{E}_n \left| f_L(\mathbf{x}; \mathbf{W}) + f_Q(\mathbf{x}; \mathbf{W}) - f(\mathbf{x}; \mathbf{W}) \right| \\ \leq \ & \max_{i \in [n]} \langle \nabla_{\mathbf{W}} f(\mathbf{x}_i; \mathbf{W}), \tilde{\mathbf{W}} \rangle^2 \cdot C d^{3/2} m^{-1/4} \|\mathbf{W}\|_{2,4}^3.\end{aligned}$$

Also, by Lemma 26 we can bound

$$
\begin{aligned}
&\mathbb{E}_n \left| \left(\langle \nabla_{\mathbf{W}} f_L(\mathbf{x}; \mathbf{W}) + \nabla_{\mathbf{W}} f_Q(\mathbf{x}; \mathbf{W}), \tilde{\mathbf{W}} \rangle \right)^2 - \langle \nabla_{\mathbf{W}} f(\mathbf{x}; \mathbf{W}), \tilde{\mathbf{W}} \rangle^2 \right| \\
\leq\ & \mathbb{E}_n \left[\left(\frac{1}{\sqrt{m}} \sum_{r=1}^m |\tilde{\mathbf{w}}_r^T \mathbf{x}| |\mathbf{w}_r^T \mathbf{x}|^2 \right) \cdot \left| \langle \nabla_{\mathbf{W}} f_L(\mathbf{x}; \mathbf{W}) + \nabla_{\mathbf{W}} f_Q(\mathbf{x}; \mathbf{W}), \tilde{\mathbf{W}} \rangle + \langle \nabla_{\mathbf{W}} f(\mathbf{x}; \mathbf{W}), \tilde{\mathbf{W}} \rangle \right| \right] \\
\leq\ & \mathbb{E}_n \left[\left(\frac{1}{\sqrt{m}} \sum_{r=1}^m |\tilde{\mathbf{w}}_r^T \mathbf{x}| |\mathbf{w}_r^T \mathbf{x}|^2 \right) \cdot \left(\frac{1}{\sqrt{m}} \sum_{r=1}^m |\tilde{\mathbf{w}}_r^T \mathbf{x}| |\mathbf{w}_r^T \mathbf{x}|^2 + 2 \langle \nabla f(\mathbf{x}; \mathbf{W}), \tilde{\mathbf{W}} \rangle \right) \right] \\
\lesssim\ & \mathbb{E}_n \left[\frac{1}{m} \left(\sum_{r=1}^m |\tilde{\mathbf{w}}_r^T \mathbf{x}| |\mathbf{w}_r^T \mathbf{x}|^2 \right)^2 \right] + \max_{i \in [n]} \langle \nabla f(\mathbf{x}_i, \mathbf{W}), \tilde{\mathbf{W}} \rangle \cdot \mathbb{E}_n \left[\frac{1}{\sqrt{m}} \sum_{r=1}^m |\tilde{\mathbf{w}}_r^T \mathbf{x}| |\mathbf{w}_r^T \mathbf{x}|^2 \right] \\
\lesssim\ & d^3 \sum_{r=1}^m \|\tilde{\mathbf{w}}_r\|^2 \|\mathbf{w}_r\|^4 + \max_{i \in [n]} \langle \nabla f(\mathbf{x}_i, \mathbf{W}), \tilde{\mathbf{W}} \rangle \cdot d^{3/2} m^{-\frac{1}{4}} \left(\|\mathbf{W}\|_{2,4}^3 + \|\tilde{\mathbf{W}}\|_{2,4}^3 \right) \\
\lesssim\ & d^3 \|\mathbf{W}\|_{2,4}^4 \|\tilde{\mathbf{W}}\|_{2,\infty}^2 + \max_{i \in [n]} \langle \nabla f(\mathbf{x}_i, \mathbf{W}), \tilde{\mathbf{W}} \rangle \cdot d^{3/2} m^{-\frac{1}{4}} \left(\|\mathbf{W}\|_{2,4}^3 + \|\tilde{\mathbf{W}}\|_{2,4}^3 \right).
\end{aligned}
$$

□

C.2 Proof of Lemma 3

Proof. We would first like to show that the quadratic model has good landscape properties. To prove this, we begin by showing any approximate stationary point must be "localized," in that the values of the regularizers at these stationary points must not be too large.

Lemma 28 (Localization)**.** *Let* $\lambda_2 = \varepsilon_{min}$, $\lambda_3 = m^{\frac{1}{2}} d^{-\frac{k-1}{2}} \varepsilon_{min}$, $\lambda_4 = d^{-2(k-1)} \varepsilon_{min}$. *Assume* $m \geq \max\left(d^{4k+4} n^2 \varepsilon_{min}^{-2}, d^{16(k+1)/3} \varepsilon_{min}^{-8/3}, n^4 \varepsilon_{min}^{-4} \right)$, *and* $\nu \leq m^{-\frac{1}{4}}$. *Then, for any* ν*-first-order stationary point* $\mathbf{W}$ *of* L_λ, *we have*

$$
\begin{aligned}
\mathcal{R}_2(\mathbf{W}) &\leq d^{2(k+1)/3} \varepsilon_{min}^{-4/3} \\
\mathcal{R}_3(\mathbf{W}) &\leq m^{-\frac{1}{2}} d^{\frac{7k+1}{6}} \varepsilon_{min}^{-4/3} \\
\mathcal{R}_4(\mathbf{W}) &\leq d^{\frac{8k-4}{3}} \varepsilon_{min}^{-4/3}.
\end{aligned}
$$

Next, we show that for these localized points, the landscape of $\hat{L}^Q$ is "good."

Lemma 29. *Let* $\mathbf{W}_L^* = \mathbf{P}_{\leq k} \mathbf{W}^*$, $\mathbf{W}_L = \mathbf{P}_{\leq k} \mathbf{W}$. *There exists a universal constant* C *such that*

$$
\mathbb{E}_{\mathbf{S}} \left[\nabla^2 \hat{L}^Q(\mathbf{W})[\mathbf{W}^* \mathbf{S}, \mathbf{W}^* \mathbf{S}] \right] - \langle \nabla \hat{L}^Q(\mathbf{W}), \mathbf{W} - 2\mathbf{W}_L^* + \mathbf{W}_L \rangle + 2\hat{L}^Q(\mathbf{W}) - 2\hat{L}^Q(\mathbf{W}^*) \tag{49}
$$

$$
\leq C \left(m^{-\frac{1}{2}} d^{\frac{k+1}{2}} + m^{-1} d^{\frac{k+3}{2}} \|\mathbf{W}\|_{2,4}^2 + d^2 m^{-\frac{1}{2}} \|\mathbf{W}\|_F (\|\mathbf{W}_L^*\|_F + \|\mathbf{W}_L\|_F) + \mathcal{R}_3(\mathbf{W}^*)^{\frac{1}{2}} + \mathcal{R}_3(\mathbf{W})^{\frac{1}{2}} \right) \tag{50}
$$

As a corollary, we obtain that for localized $\mathbf{W}$, this error term can be made arbitrarily small for sufficiently large m.

Corollary 5. *Assume* $m \gtrsim d^{\frac{14k+20}{3}} \varepsilon_{min}^{-22/3}$. *Then, additionally under the assumptions of Lemma 3, any* ν*-first-order stationary point* $\mathbf{W}$ *satisfies*

$$
\mathbb{E}_{\mathbf{S}} \left[\nabla^2 \hat{L}^Q(\mathbf{W})[\mathbf{W}^* \mathbf{S}, \mathbf{W}^* \mathbf{S}] \right] - \langle \nabla \hat{L}^Q(\mathbf{W}), \mathbf{W} - 2\mathbf{W}_L^* + \mathbf{W}_L \rangle + 2\hat{L}^Q(\mathbf{W}) - 2\hat{L}^Q(\mathbf{W}^*) \leq \varepsilon_{min}.
$$

Our final step is to show that the landscape of L_λ is good. To do this, we first show that the landscapes of the quadratic model and original model couple for localized points. This coupling is given by the following lemma.

Lemma 30. *Assume the conditions of Lemma 3, Corollary 5, and let* $\mathbf{W}$ *be a* ν*-first order stationary point. Furthermore, assume* $m \gtrsim n^4 d^{\frac{26(k+1)}{3}} \varepsilon_{min}^{-22/3}$*. Then,*

$$\begin{aligned}
\left|\hat{L}(\mathbf{W}) - \hat{L}^Q(\mathbf{W})\right| &\leq \varepsilon_{min} \\
\left|\hat{L}(\mathbf{W}^*) - \hat{L}^Q(\mathbf{W}^*)\right| &\leq \varepsilon_{min} \\
\left|\langle \nabla \hat{L}(\mathbf{W}), \mathbf{W} - 2\mathbf{W}_L^* + \mathbf{W}_L\rangle - \langle \nabla \hat{L}^Q(\mathbf{W}), \mathbf{W} - 2\mathbf{W}_L^* + \mathbf{W}_L\rangle\right| &\leq \varepsilon_{min} \\
\left|\mathbb{E}_{\mathbf{S}}\left[\nabla^2 \hat{L}(\mathbf{W})[\mathbf{W}^*\mathbf{S}, \mathbf{W}^*\mathbf{S}]\right] - \mathbb{E}_{\mathbf{S}}\left[\nabla^2 \hat{L}^Q(\mathbf{W})[\mathbf{W}^*\mathbf{S}, \mathbf{W}^*\mathbf{S}]\right]\right| &\leq \varepsilon_{min}.
\end{aligned}$$

An immediate corollary is that the landscape of $\hat{L}$ must be good.

Corollary 6. *Let* $\mathbf{W}$ *be a* ν*-first-order stationary point of* L_λ*. Under the same conditions of Lemma 3, Corollary 5, Lemma 30, we have*

$$\mathbb{E}_{\mathbf{S}}\left[\nabla^2 \hat{L}(\mathbf{W})[\mathbf{W}^*\mathbf{S}, \mathbf{W}^*\mathbf{S}]\right] - \langle \nabla \hat{L}(\mathbf{W}), \mathbf{W} - 2\mathbf{W}_L^* + \mathbf{W}_L\rangle + 2\hat{L}(\mathbf{W}) - 2\hat{L}(\mathbf{W}^*) \lesssim \varepsilon_{min}.$$

To conclude, we must show that adding the regularizers has a benign effect on the landscape

Lemma 31. *Define*

$$\mathcal{R}(\mathbf{W}) = \lambda_1 \mathcal{R}_1(\mathbf{W}) + \lambda_2 \mathcal{R}_2(\mathbf{W}) + \lambda_3 \mathcal{R}_3(\mathbf{W}) + \lambda_4 \mathcal{R}_4(\mathbf{W}) \tag{51}$$

to be the total regularization term. Under the conditions of Lemma 3, we have

$$\mathbb{E}_{\mathbf{S}} \nabla^2 \mathcal{R}(\mathbf{W})[\mathbf{W}^*\mathbf{S}, \mathbf{W}^*\mathbf{S}] - \langle \nabla \mathcal{R}(\mathbf{W}), \mathbf{W} - 2\mathbf{W}_L^* + \mathbf{W}_L\rangle + 2\mathcal{R}(\mathbf{W}) - 2\mathcal{R}(\mathbf{W}^*) \lesssim \varepsilon_{min}. \tag{52}$$

Lemma 3 now directly follows by adding the results of Corollary 6 and Lemma 31.

□

C.3 Proof of Corollary 1

Proof. Let $\mathbf{W}$ be an (ν, γ)-SOSP of $L_\lambda(\mathbf{W})$. Then

$$\begin{aligned}
\langle \nabla L_\lambda(\mathbf{W}), \mathbf{W} - 2\mathbf{W}_L^* + \mathbf{W}_L\rangle &\leq \nu \|\mathbf{W} - 2\mathbf{W}_L^* + \mathbf{W}_L\|_F \\
&\leq \nu \cdot m^{1/4} d^{\frac{k}{3} - \frac{1}{6}} \varepsilon^{-1/6} \\
&\leq m^{-1/4} d^{\frac{k}{3} - \frac{1}{6}} \varepsilon^{-1/6} \\
&\leq \varepsilon_{min},
\end{aligned}$$

since we have chosen $\nu \leq m^{-1/2}, m \geq d^{\frac{4k-2}{3}} \varepsilon^{-14/3}$. Also,

$$\begin{aligned}
\nabla^2 L_\lambda(\mathbf{W})[\mathbf{W}^*\mathbf{S}, \mathbf{W}^*\mathbf{S}] &\geq -\gamma \|\mathbf{W}^*\|_F^2 \\
&\geq -\gamma m^{1/2} d^{(k-1)/2} \\
&\geq -m^{-1/4} d^{(k-1)/2} \\
&\geq -\varepsilon_{min},
\end{aligned}$$

since we have chosen $\gamma \leq m^{-3/4}, m \geq d^{2(k-1)} \varepsilon^{-4}$. Altogether, we can bound

$$L_\lambda(\mathbf{W}) \lesssim L_\lambda(\mathbf{W}^*) + \langle \nabla L_\lambda(\mathbf{W}), \mathbf{W} - 2\mathbf{W}_L^* + \mathbf{W}_L\rangle - \mathbb{E}_{\mathbf{S}}[\nabla^2 L_\lambda(\mathbf{W})[\mathbf{W}^*\mathbf{S}, \mathbf{W}^*\mathbf{S}]] + \varepsilon_{min} \lesssim \varepsilon_{min},$$

as desired. □

C.4 Proofs of Intermediate Results

C.4.1 Proof of Lemma 28

Proof. Let $\mathbf{W}$ be an ν-first-order stationary point of L_λ. Then

$$\langle \nabla L_\lambda(\mathbf{W}), \mathbf{W} \rangle \leq \nu \|\mathbf{W}\|_F. \tag{53}$$

We have that

$$\langle \nabla \hat{L}^Q(\mathbf{W}), \mathbf{W} \rangle = \mathbb{E}_n \left[\ell'(y, f_L(\mathbf{x}; \mathbf{W}) + f_Q(\mathbf{x}; \mathbf{W})) \cdot (2 f_Q(\mathbf{x}; \mathbf{W}) + f_L(\mathbf{x}; \mathbf{W})) \right].$$

First, by convexity we can bound

$$\mathbb{E}_n \left[\ell'(y, f_L(\mathbf{x}; \mathbf{W}) + f_Q(\mathbf{x}; \mathbf{W})) \cdot (f_Q(\mathbf{x}; \mathbf{W}) + f_L(\mathbf{x}; \mathbf{W})) \right] \geq \hat{L}^Q(\mathbf{W}) - \hat{L}^Q(\mathbf{0}) \geq -1.$$

Secondly, we can bound

$$\left| \mathbb{E}_n \left[\ell'(y, f_L(\mathbf{x}; \mathbf{W}) + f_Q(\mathbf{x}; \mathbf{W})) f_Q(\mathbf{x}; \mathbf{W}) \right] \right| \leq \mathbb{E}_n \left[\frac{1}{\sqrt{m}} \sum_{r=1}^m (\mathbf{w}_r^T \mathbf{x})^2 \right] \leq \frac{d}{\sqrt{m}} \|\mathbf{W}\|_F^2.$$

Finally, by Lemma 23,

$$\left| \langle \nabla \hat{L}^Q(\mathbf{W}), \mathbf{W} \rangle - \langle \nabla \hat{L}(\mathbf{W}), \mathbf{W} \rangle \right| \lesssim d^{3/2} m^{-1/4} \|\mathbf{W}\|_{2,4}^3 \cdot \max_{i \in [n]} \left| \langle \nabla f(\mathbf{x}; \mathbf{W}), \mathbf{W} \rangle \right|.$$

Altogether,

$$\langle \nabla \hat{L}(\mathbf{W}), \mathbf{W} \rangle \geq -1 - \frac{d}{\sqrt{m}} \|\mathbf{W}\|_F^2 - C d^{3/2} m^{-1/4} \|\mathbf{W}\|_{2,4}^3 \cdot \max_{i \in [n]} \left| \langle \nabla f(\mathbf{x}; \mathbf{W}), \mathbf{W} \rangle \right|.$$

We next turn to the regularizers. $\mathcal{R}_i$ for $i = 1, 2, 3$ are all quadratics, so $\langle \nabla \mathcal{R}_i(\mathbf{W}), \mathbf{W} \rangle = 2\mathcal{R}_i(\mathbf{W})$. Also it is true that $\langle \nabla \mathcal{R}_4(\mathbf{W}), \mathbf{W} \rangle = 8\mathcal{R}_4(\mathbf{W})$. Plugging into (53), we get

$$\begin{aligned}
\nu \|\mathbf{W}\|_F &\geq \left\langle \nabla \left(\hat{L}(\mathbf{W}) + \lambda_1 \mathcal{R}_1(\mathbf{W}) + \lambda_2 \mathcal{R}_2(\mathbf{W}) + \lambda_3 \mathcal{R}_3(\mathbf{W}) + \lambda_4 \mathcal{R}_4(\mathbf{W}) \right), \mathbf{W} \right\rangle \\
&\geq 2\lambda_1 \mathcal{R}_1(\mathbf{W}) + 2\lambda_2 \mathcal{R}_2(\mathbf{W}) + 2\lambda_3 \mathcal{R}_3(\mathbf{W}) + 8\lambda_4 \mathcal{R}_4(\mathbf{W}) \\
&\quad - 1 - d m^{-\frac{1}{2}} \|\mathbf{W}\|_F^2 - C d^{3/2} m^{-\frac{1}{4}} \|\mathbf{W}\|_{2,4}^3 \cdot \max_{i \in [n]} \left| \langle \nabla f(\mathbf{x}; \mathbf{W}), \mathbf{W} \rangle \right|.
\end{aligned}$$

Therefore (using $\nu \leq m^{-1/4}$),

$$\begin{aligned}
&\lambda_1 \mathcal{R}_1(\mathbf{W}) + \lambda_2 \mathcal{R}_2(\mathbf{W}) + \lambda_3 \mathcal{R}_3(\mathbf{W}) + \lambda_4 \mathcal{R}_4(\mathbf{W}) \\
\lesssim\ & 1 + d m^{-\frac{1}{2}} \|\mathbf{W}\|_F^2 + d^{3/2} m^{-\frac{1}{4}} \|\mathbf{W}\|_{2,4}^3 \cdot \max_{i \in [n]} \left| \langle \nabla f(\mathbf{x}; \mathbf{W}), \mathbf{W} \rangle \right| + \nu \|\mathbf{W}\|_F \\
\lesssim\ & 1 + d m^{-\frac{1}{2}} \|\mathbf{W}\|_F^2 + d^{3/2} m^{-\frac{1}{4}} \|\mathbf{W}\|_{2,4}^3 \cdot (n^{\frac{1}{2}} \mathbb{E}_n \left[(f_L(\mathbf{x}; \mathbf{W}))^2 \right]^{\frac{1}{2}} + d n^{\frac{1}{2}} m^{-\frac{1}{2}} \|\mathbf{W}\|_F^2),
\end{aligned}$$

where the last step follows from Lemma 27.

We can bound

$$\begin{aligned}
\mathbb{E}_n \left[(f_L(\mathbf{x}; \mathbf{W}))^2 \right] &= \mathbb{E}_n \left[(f_L(\mathbf{x}; \mathbf{P}_{>k} \mathbf{W}) + f_L(\mathbf{x}; \mathbf{P}_{\leq k} \mathbf{W}))^2 \right] \\
&\lesssim \mathbb{E}_n \left[(f_L(\mathbf{x}; \mathbf{P}_{>k} \mathbf{W}))^2 + (f_L(\mathbf{x}; \mathbf{P}_{\leq k} \mathbf{W}))^2 \right] \\
&\lesssim \mathbb{E}_n \left[(f_L(\mathbf{x}; \mathbf{P}_{>k} \mathbf{W}))^2 \right] + \|f_L(\mathbf{x}; \mathbf{P}_{\leq k} \mathbf{W})\|_{L^2}^2 \\
&= \mathcal{R}_2(\mathbf{W}) + \mathcal{R}_3(\mathbf{W}).
\end{aligned}$$

Therefore by AM-GM,

$$\lambda_2 \mathcal{R}_2(\mathbf{W}) + \lambda_3 \mathcal{R}_3(\mathbf{W}) + \lambda_4 \mathcal{R}_4(\mathbf{W}) \tag{54}$$

$$\lesssim 1 + d m^{-\frac{1}{2}} \|\mathbf{W}\|_F^2 + m^{-\frac{1}{4}} d^3 \|\mathbf{W}\|_{2,4}^6 + m^{-\frac{1}{4}} n \mathbb{E}_n \left[(f_L(\mathbf{x}; \mathbf{W}))^2 \right] + m^{-\frac{1}{4}} \|\mathbf{W}\|_{2,4}^3 d^{5/2} n^{\frac{1}{2}} m^{-\frac{1}{2}} \|\mathbf{W}\|_F^2 \tag{55}$$

$$\lesssim 1 + d \|\mathbf{W}\|_{2,4}^2 + m^{-\frac{1}{4}} d^{5/2} n^{\frac{1}{2}} \|\mathbf{W}\|_{2,4}^5 + m^{-\frac{1}{4}} d^3 \|\mathbf{W}\|_{2,4}^6 + m^{-\frac{1}{4}} n \left(\mathcal{R}_2(\mathbf{W}) + \mathcal{R}_3(\mathbf{W}) \right), \tag{56}$$

where the last step uses $\|\mathbf{W}\|_F \le m^{\frac{1}{4}}\|\mathbf{W}\|_{2,4}$.

Since $m^{-\frac{1}{4}}n \le \varepsilon_{min}/2 \le \lambda_2/2, \lambda_3/2$, we have

$$\lambda_4\|\mathbf{W}\|_{2,4}^8 \lesssim 1 + d\|\mathbf{W}\|_{2,4}^2 + m^{-\frac{1}{4}}d^{5/2}n^{\frac{1}{2}}\|\mathbf{W}\|_{2,4}^5 + m^{-\frac{1}{4}}d^3\|\mathbf{W}\|_{2,4}^6 \tag{57}$$

Therefore, plugging in $\lambda_4 = d^{-2(k-1)}\varepsilon_{min}$,

$$\|\mathbf{W}\|_{2,4} \lesssim \max\left(d^{\frac{k-1}{4}}\varepsilon_{min}^{-1/8}, d^{\frac{2k-1}{6}}\varepsilon_{min}^{-1/6}, m^{-1/12}d^{2k/3+1/6}n^{1/6}\varepsilon_{min}^{-1/3}, m^{-\frac{1}{8}}d^{k+1/2}\varepsilon_{min}^{-1/2}\right). \tag{58}$$

Since $\varepsilon_{min} < 1$ we trivially have $d^{\frac{k-1}{4}}\varepsilon_{min}^{-1/8} < d^{\frac{2k-1}{6}}\varepsilon_{min}^{-1/6}$. Also, since $m \ge d^{4k+4}n^2\varepsilon_{min}^{-2}$, we have

$$d^{\frac{2k-1}{6}}\varepsilon_{min}^{-1/6} \ge m^{-1/12}d^{\frac{2k}{3}+\frac{1}{6}}n^{1/6}\varepsilon_{min}^{-1/3}.$$

Additionally, assuming $m \ge d^{16/3(k+1)}\varepsilon_{min}^{-8/3}$, we have

$$d^{\frac{2k-1}{6}}\varepsilon_{min}^{-1/6} \ge m^{-\frac{1}{8}}d^{k+1/2}\varepsilon_{min}^{-1/2}.$$

Therefore we can bound

$$\|\mathbf{W}\|_{2,4} \lesssim d^{\frac{2k-1}{6}}\varepsilon_{min}^{-1/6},$$

and thus

$$\mathcal{R}_4(\mathbf{W}) = \|\mathbf{W}\|_{2,4}^8 \lesssim d^{\frac{8k-4}{3}}\varepsilon_{min}^{-4/3}. \tag{59}$$

In this case, the RHS of (57) can be upper bounded by $d\|\mathbf{W}\|_{2,4}^2$. Plugging back into (56), we get

$$\lambda_2\mathcal{R}_2(\mathbf{W}) + \lambda_3\mathcal{R}_3(\mathbf{W}) \lesssim d^{2(k+1)/3}\varepsilon_{min}^{-1/3},$$

which yields the bounds

$$\mathcal{R}_2(\mathbf{W}) \lesssim d^{2(k+1)/3}\varepsilon_{min}^{-4/3} \tag{60}$$

$$\mathcal{R}_3(\mathbf{W}) \lesssim m^{-\frac{1}{2}}d^{\frac{7k+1}{6}}\varepsilon_{min}^{-4/3}. \tag{61}$$

□

C.4.2 Proof of Lemma 29

Proof. Observe that

$$\begin{aligned}
&\nabla\hat{L}^Q(\mathbf{W})[\tilde{\mathbf{W}}, \tilde{\mathbf{W}}] \\
&= \mathbb{E}_n\left[\ell'(y, f_L(\mathbf{x};\mathbf{W}) + f_Q(\mathbf{x};\mathbf{W}))\frac{1}{\sqrt{m}}\sum_{r=1}^m a_r\sigma''(\mathbf{w}_{0,r}^T\mathbf{x})(\tilde{\mathbf{w}}_r^T\mathbf{x})^2\right] \\
&+ \mathbb{E}_n\left[\ell''(y, f_L(\mathbf{x};\mathbf{W}) + f_Q(\mathbf{x};\mathbf{W}))\left(\langle\nabla_{\mathbf{W}}f_L(\mathbf{x};\mathbf{W}) + \nabla_{\mathbf{W}}f_Q(\mathbf{x};\mathbf{W}), \tilde{\mathbf{W}}\rangle\right)^2\right],
\end{aligned}$$

and therefore for any diagonal matrix of random signs $\mathbf{S} = \mathrm{diag}(\sigma_1, \ldots, \sigma_m)$,

$$\begin{aligned}
&\nabla\hat{L}^Q(\mathbf{W})[\mathbf{W}^*\mathbf{S}, \mathbf{W}^*\mathbf{S}] \\
&= 2\mathbb{E}_n\left[\ell'(y, f_L(\mathbf{x};\mathbf{W}) + f_Q(\mathbf{x};\mathbf{W}))f_Q(\mathbf{x};\mathbf{W}^*)\right] \\
&+ \mathbb{E}_n\left[\ell''(y, f_L(\mathbf{x};\mathbf{W}) + f_Q(\mathbf{x};\mathbf{W}))\left(\frac{1}{\sqrt{m}}\sum_{r=1}^m a_r\sigma_r\left(\sigma'(\mathbf{w}_{0,r}^T\mathbf{x})\mathbf{x}^T\mathbf{w}_r^* + \sigma''(\mathbf{w}_{0,r}^T\mathbf{x})\mathbf{w}_r^T\mathbf{x}\mathbf{x}^T\mathbf{w}_r^*\right)\right)^2\right],
\end{aligned}$$

The expectation of the second term over the random signs $\mathbf{S}$ can be upper bounded by

$$\begin{aligned}
&\mathbb{E}_{\mathbf{S}}\mathbb{E}_n\left[\left(\frac{1}{\sqrt{m}}\sum_{r=1}^m a_r\sigma_r\left(\sigma'(\mathbf{w}_{0,r}^T\mathbf{x})\mathbf{x}^T\mathbf{w}_r^* + \sigma''(\mathbf{w}_{0,r}^T\mathbf{x})\mathbf{w}_r^T\mathbf{x}\mathbf{x}^T\mathbf{w}_r^*\right)\right)^2\right] \\
=\ &\mathbb{E}_n\left[\frac{1}{m}\sum_{r=1}^m\left(\sigma'(\mathbf{w}_{0,r}^T\mathbf{x})\mathbf{x}^T\mathbf{w}_r^* + \sigma''(\mathbf{w}_{0,r}^T\mathbf{x})\mathbf{w}_r^T\mathbf{x}\mathbf{x}^T\mathbf{w}_r^*\right)^2\right] \\
\lesssim\ &\mathbb{E}_n\left[\frac{1}{m}\sum_{r=1}^m(\mathbf{x}^T\mathbf{w}_r^*)^2 + (\mathbf{w}_r^T\mathbf{x}\mathbf{x}^T\mathbf{w}_r^*)^2\right] \\
\lesssim\ &\frac{1}{m}\sum_{r=1}^m\left(d\|\mathbf{w}_r^*\|^2 + d^2\|\mathbf{w}_r\|^2\|\mathbf{w}_r^*\|^2\right) \\
\lesssim\ &\frac{d}{m}\|\mathbf{W}^*\|_F^2 + \frac{d^2}{m}\|\mathbf{W}\|_{2,4}^2\|\mathbf{W}^*\|_{2,4}^2 \\
\lesssim\ &m^{-\frac{1}{2}}d^{\frac{k+1}{2}} + m^{-1}d^{\frac{k+3}{2}}\|\mathbf{W}\|_{2,4}^2
\end{aligned}$$

Therefore

$$\mathbb{E}_{\mathbf{S}}\left[\nabla^2\hat{L}^Q(\mathbf{W})[\mathbf{W}^*\mathbf{S},\mathbf{W}^*\mathbf{S}]\right] \tag{62}$$

$$\leq 2\mathbb{E}_n\left[\ell'(y, f_L(\mathbf{x};\mathbf{W}) + f_Q(\mathbf{x};\mathbf{W}))f_Q(\mathbf{x};\mathbf{W}^*)\right] + C(m^{-\frac{1}{2}}d^{\frac{k+1}{2}} + m^{-1}d^{\frac{k+3}{2}}\|\mathbf{W}\|_{2,4}^2) \tag{63}$$

Next, define $\mathbf{\Delta} = \mathbf{W} - 2\mathbf{W}_L^* + \mathbf{W}_L$. We have

$$\begin{aligned}
&\langle\hat{L}^Q(\mathbf{W}), \mathbf{\Delta}\rangle \\
&= \mathbb{E}_n\left[\ell'(y, f_L(\mathbf{x};\mathbf{W}) + f_Q(\mathbf{x};\mathbf{W}))\left(\frac{1}{\sqrt{m}}\sum_{r=1}^m\left(a_r\sigma'(\mathbf{w}_{0,r}^T\mathbf{x})\mathbf{x}^T\mathbf{\Delta}_r + a_r\sigma''(\mathbf{w}_{0,r}^T\mathbf{x})\mathbf{w}_r^T\mathbf{x}\mathbf{x}^T\mathbf{\Delta}_r\right)\right)\right] \\
&= \mathbb{E}_n\left[\ell'(y, f_L(\mathbf{x};\mathbf{W}) + f_Q(\mathbf{x};\mathbf{W}))\left(f_L(\mathbf{x};\mathbf{\Delta}) + \frac{1}{\sqrt{m}}\sum_{r=1}^m a_r\sigma''(\mathbf{w}_{0,r}^T\mathbf{x})\mathbf{w}_r^T\mathbf{x}\mathbf{x}^T\mathbf{\Delta}_r\right)\right] \\
&= \mathbb{E}_n\left[\ell'(y, f_L(\mathbf{x};\mathbf{W}) + f_Q(\mathbf{x};\mathbf{W}))\left(f_L(\mathbf{x};\mathbf{W}) - 2f_L(\mathbf{x};\mathbf{W}_L^*) + f_L(\mathbf{x};\mathbf{W}_L) + 2f_Q(\mathbf{x};\mathbf{W})\right)\right] \\
&\quad + \mathbb{E}_n\left[\ell'(y, f_L(\mathbf{x};\mathbf{W}) + f_Q(\mathbf{x};\mathbf{W}))\left(\frac{1}{\sqrt{m}}\sum_{r=1}^m a_r\sigma''(\mathbf{w}_{0,r}^T\mathbf{x})\mathbf{w}_r^T\mathbf{x}\mathbf{x}^T(-2(\mathbf{w}_L^*)_r + (\mathbf{w}_L)_r)\right)\right].
\end{aligned}$$

The second term can be bounded in magnitude as

$$\begin{aligned}
&\left|\mathbb{E}_n\left[\ell'(y, f_L(\mathbf{x};\mathbf{W}) + f_Q(\mathbf{x};\mathbf{W}))\left(\frac{1}{\sqrt{m}}\sum_{r=1}^m a_r\sigma''(\mathbf{w}_{0,r}^T\mathbf{x})\mathbf{w}_r^T\mathbf{x}\mathbf{x}^T(-2(\mathbf{w}_L^*)_r + (\mathbf{w}_L)_r)\right)\right]\right| \\
\leq\ &\mathbb{E}_n\left[\frac{1}{\sqrt{m}}\sum_{r=1}^m\left|\mathbf{w}_r^T\mathbf{x}\mathbf{x}^T(-2(\mathbf{w}_L^*)_r + (\mathbf{w}_L)_r)\right|\right] \\
\lesssim\ &\frac{d^2}{\sqrt{m}}\sum_{r=1}^m\|\mathbf{w}_r\|(\|(\mathbf{w}_L^*)_r\| + \|(\mathbf{w}_L)_r\|) \\
\leq\ &d^2m^{-\frac{1}{2}}\|\mathbf{W}\|_F(\|\mathbf{W}_L^*\|_F + \|\mathbf{W}_L\|_F)
\end{aligned}$$

Also, we have that

$$\begin{aligned}
|\mathbb{E}_n[\ell'(y, f_L(\mathbf{x};\mathbf{W}) + f_Q(\mathbf{x};\mathbf{W}))(f_L(\mathbf{x};\mathbf{W}) - f_L(\mathbf{x};\mathbf{W}_L))]| &\leq \mathbb{E}_n\left|f_L(\mathbf{x};\mathbf{W} - \mathbf{W}_L)\right| \\
&\leq \mathbb{E}_n\left[(f_L(\mathbf{x};\mathbf{P}_{>k}\mathbf{W}))^2\right]^{\frac{1}{2}} \\
&= \mathcal{R}_3(\mathbf{W})^{\frac{1}{2}},
\end{aligned}$$

and similarly

$$|\mathbb{E}_n[\ell'(y, f_L(\mathbf{x};\mathbf{W}) + f_Q(\mathbf{x};\mathbf{W}))(f_L(\mathbf{x};\mathbf{W}^*) - f_L(\mathbf{x};\mathbf{W}_L^*))]| \leq \mathbb{E}_n\left|f_L(\mathbf{x};\mathbf{W}^* - \mathbf{W}_L^*)\right| \lesssim \mathcal{R}_3(\mathbf{W}^*)^{\frac{1}{2}}.$$

Altogether, we have for some constant $C' > 0$,

$$\langle \hat{L}^Q(\mathbf{W}), \mathbf{\Delta} \rangle \geq 2\mathbb{E}_n \left[\ell'(y, f_L(\mathbf{x}; \mathbf{W}) + f_Q(\mathbf{x}; \mathbf{W}))(f_Q(\mathbf{x}; \mathbf{W}) + f_L(\mathbf{x}; \mathbf{W}) - f_L(\mathbf{x}; \mathbf{W}^*)) \right] \tag{64}$$

$$- C' \cdot (d^2 m^{-\frac{1}{2}} \|\mathbf{W}\|_F (\|\mathbf{W}_L^*\|_F + \|\mathbf{W}_L\|_F) + \mathcal{R}_3(\mathbf{W}^*)^{\frac{1}{2}} + \mathcal{R}_3(\mathbf{W})^{\frac{1}{2}}). \tag{65}$$

Finally, by convexity of ℓ we have

$$\hat{L}^Q(\mathbf{W}) - \hat{L}^Q(\mathbf{W}^*) \tag{66}$$

$$\leq \mathbb{E}_n \left[\ell'(y, f_L(\mathbf{x}; \mathbf{W}) + f_Q(\mathbf{x}; \mathbf{W}))(f_L(\mathbf{x}; \mathbf{W}) + f_Q(\mathbf{x}; \mathbf{W}) - f_L(\mathbf{x}; \mathbf{W}^*) - f_Q(\mathbf{x}; \mathbf{W}^*)) \right]. \tag{67}$$

Combining (63), (65), and (67), we get that

$$\mathbb{E}_{\mathbf{S}} \left[\nabla^2 \hat{L}^Q(\mathbf{W})[\mathbf{W}^*\mathbf{S}, \mathbf{W}^*\mathbf{S}] \right] - \langle \hat{L}^Q(\mathbf{W}), \mathbf{\Delta} \rangle + 2(\hat{L}^Q(\mathbf{W}) - \hat{L}^Q(\mathbf{W}^*))$$

$$\leq C \left(m^{-\frac{1}{2}} d^{\frac{k+1}{2}} + m^{-1} d^{\frac{k+3}{2}} \|\mathbf{W}\|_{2,4}^2 + d^2 m^{-\frac{1}{2}} \|\mathbf{W}\|_F (\|\mathbf{W}_L^*\|_F + \|\mathbf{W}_L\|_F) + \mathcal{R}_3(\mathbf{W}^*)^{\frac{1}{2}} + \mathcal{R}_3(\mathbf{W})^{\frac{1}{2}} \right),$$

as desired. □

C.4.3 Proof of Corollary 5

Proof. First, note that we have the bounds

$$\|\mathbf{W}_L^*\|_F \lesssim d^{(k-1)/2}$$
$$\mathcal{R}_3(\mathbf{W}^*)^{\frac{1}{2}} \lesssim m^{-\frac{1}{4}} d^{(k-1)/4}.$$

Also, by Lemma 28, for ν-first order stationary point $\mathbf{W}$ we can bound

$$\|\mathbf{W}\|_F \leq m^{1/4} \|\mathbf{W}\|_{2,4} \leq m^{1/4} d^{\frac{k}{3} - \frac{1}{6}} \varepsilon_{min}^{-1/6}$$
$$\mathcal{R}_3(\mathbf{W})^{\frac{1}{2}} \leq m^{-1/4} d^{\frac{7k+1}{12}} \varepsilon_{min}^{-2/3}.$$

Furthermore, since $\mathrm{vec}(\mathbf{W}_L) \in \mathrm{span}(\mathbf{P}_{\leq k})$, we can write

$$\mathcal{R}_2(\mathbf{W}_L) = \mathrm{vec}(\mathbf{W}_L)^T \mathbf{\Sigma}_{\leq k} \mathrm{vec}(\mathbf{W}_L) \geq \lambda_{n_k}(\mathbf{\Sigma}) \|\mathbf{W}_L\|_F^2 = \Theta(d^{1-k}) \cdot \|\mathbf{W}_L\|_F^2.$$

Therefore

$$\|\mathbf{W}_L\|_F \leq d^{(k-1)/2} \mathcal{R}_2(\mathbf{W})^{\frac{1}{2}} \leq d^{\frac{5k-1}{6}} \varepsilon_{min}^{-2/3}.$$

Altogether, by Lemma 29, we can bound

$$\mathbb{E}_{\mathbf{S}} \left[\nabla^2 \hat{L}^Q(\mathbf{W})[\mathbf{W}^*\mathbf{S}, \mathbf{W}^*\mathbf{S}] \right] - \langle \nabla \hat{L}^Q(\mathbf{W}), \mathbf{W} - 2\mathbf{W}_L^* + \mathbf{W}_L \rangle + 2\hat{L}^Q(\mathbf{W}) - 2\hat{L}^Q(\mathbf{W}^*)$$
$$\lesssim m^{-\frac{1}{2}} d^{\frac{k+1}{2}} + m^{-1} d^{\frac{k+3}{2}} \|\mathbf{W}\|_{2,4}^2 + d^2 m^{-\frac{1}{2}} \|\mathbf{W}\|_F (\|\mathbf{W}_L^*\|_F + \|\mathbf{W}_L\|_F) + \mathcal{R}_3(\mathbf{W}^*)^{\frac{1}{2}} + \mathcal{R}_3(\mathbf{W})^{\frac{1}{2}}$$
$$\lesssim m^{-\frac{1}{2}} d^{\frac{k+1}{2}} + m^{-1} d^{\frac{k+3}{2}} d^{\frac{2k-1}{3}} \varepsilon_{min}^{-1/3} + d^2 m^{-\frac{1}{2}} \cdot m^{\frac{1}{4}} d^{\frac{k}{3} - \frac{1}{6}} \varepsilon_{min}^{-1/6} \cdot d^{\frac{5k-1}{6}} \varepsilon_{min}^{-2/3} + m^{-\frac{1}{4}} d^{\frac{7k+1}{12}} \varepsilon_{min}^{-2/3}$$
$$= m^{-\frac{1}{2}} d^{\frac{k-1}{2}} + m^{-1} d^{\frac{7k+7}{6}} \varepsilon_{min}^{-1/3} + m^{-\frac{1}{4}} d^{\frac{7k+10}{6}} \varepsilon_{min}^{-5/6} + m^{-\frac{1}{4}} d^{\frac{7k+1}{12}} \varepsilon_{min}^{-2/3}$$
$$\leq \varepsilon_{min},$$

since we have assumed $m \gtrsim d^{\frac{14k+20}{3}} \varepsilon_{min}^{-22/3}$. □

C.4.4 Proof of Lemma 30

Loss Coupling. By Lemma 22, we can bound

$$\left| \hat{L}(\mathbf{W}) - \hat{L}^Q(\mathbf{W}) \right| \leq d^{3/2} m^{-1/4} \|\mathbf{W}\|_{2,4}^3$$
$$\lesssim m^{-1/4} d^{k+1} \varepsilon_{min}^{-1/2}$$
$$\leq \varepsilon_{min}$$

for $m \geq d^{4k+4}\varepsilon_{min}^{-6}$.

Similarly,

$$\begin{aligned}\left|\hat{L}(\mathbf{W}^*) - \hat{L}^Q(\mathbf{W}^*)\right| &\leq d^{3/2}m^{-1/4}\|\mathbf{W}^*\|_{2,4}^3 \\ &\lesssim d^{3/2}m^{-1/4}d^{3(k-1)/4} \\ &\leq \varepsilon_{min}\end{aligned}$$

since $m \geq d^{3(k+1)}\varepsilon_{min}^{-4}$.

Gradient Coupling. Next, by Lemma 23, we have

$$\begin{aligned}&\left|\langle\nabla\hat{L}(\mathbf{W}), \mathbf{W} - 2\mathbf{W}_L^* + \mathbf{W}_L\rangle - \langle\nabla\hat{L}^Q(\mathbf{W}), \mathbf{W} - 2\mathbf{W}_L^* + \mathbf{W}_L\rangle\right| \\ &\lesssim d^{3/2}m^{-1/4}(\|\mathbf{W}\|_{2,4}^3 + \|\mathbf{\Delta}\|_{2,4}^3) \cdot \max_{i\in[n]} |\langle\nabla_{\mathbf{W}} f(\mathbf{x}; \mathbf{W}), \mathbf{\Delta}\rangle|,\end{aligned}$$

where $\mathbf{\Delta} := \mathbf{W} - 2\mathbf{W}_L^* + \mathbf{W}_L$. First, observe that

$$\begin{aligned}\|\mathbf{\Delta}\|_{2,4} &\leq \|\mathbf{W}\|_{2,4} + 2\|\mathbf{W}_L^*\|_{2,4} + \|\mathbf{W}_L\|_{2,4} \\ &\leq \|\mathbf{W}\|_{2,4} + 2m^{-1/4}d^{k/2}\|\mathbf{W}_L^*\|_F + m^{-1/4}d^{k/2}\|\mathbf{W}_L\|_F,\end{aligned}$$

where we applied Lemma 21. Plugging in the bounds for $\|\mathbf{W}\|_{2,4}$, $\|\mathbf{W}_L\|_{2,4}$, we get

$$\begin{aligned}\|\mathbf{\Delta}\|_{2,4} &\leq d^{\frac{k}{3}-\frac{1}{6}}\varepsilon_{min}^{-1/6} + m^{-1/4}d^{k/2} \cdot d^{\frac{5k-1}{6}}\varepsilon_{min}^{-2/3} \\ &\leq d^{\frac{k}{3}-\frac{1}{6}}\varepsilon_{min}^{-1/6},\end{aligned}$$

since $m \geq d^{4k}\varepsilon_{min}^{-2}$.

Also, by Lemma 27,

$$\begin{aligned}&\max_{i\in[n]} |\langle\nabla_{\mathbf{W}} f(\mathbf{x}; \mathbf{W}), \mathbf{\Delta}\rangle| \\ &\leq n^{1/2}\mathbb{E}_n\left[(f_L(\mathbf{x}; \Delta))^2\right]^{1/2} + dn^{\frac{1}{2}}m^{-1/2}\|\mathbf{W}\|_F\|\mathbf{\Delta}\|_F \\ &\leq n^{1/2}\mathbb{E}_n\left[(f_L(\mathbf{x}; \Delta))^2\right]^{1/2} + dn^{\frac{1}{2}}\|\mathbf{W}\|_{2,4}\|\mathbf{\Delta}\|_{2,4}.\end{aligned}$$

Observe that

$$\begin{aligned}\mathbb{E}_n\left[(f_L(\mathbf{x}; \Delta))^2\right] &\lesssim \mathbb{E}_n\left[(f_L(\mathbf{x}; \mathbf{P}_{>k}\mathbf{W}))^2\right] + \mathbb{E}_n\left[(f_L(\mathbf{x}; \mathbf{W}_L))^2\right] + \mathbb{E}_n\left[(f_L(\mathbf{x}; \mathbf{W}_L^*))^2\right] \\ &\lesssim \mathcal{R}_3(\mathbf{W}) + \mathbb{E}_n\left[(f_L(\mathbf{x}; \mathbf{W}_L))^2\right] + \mathbb{E}_n\left[(f_L(\mathbf{x}; \mathbf{W}_L^*))^2\right],\end{aligned}$$

where we first decomposed $\mathbf{W} = \mathbf{P}_{>k}\mathbf{W} + \mathbf{W}_L$ and then used the definition of $\mathcal{R}_3$. Since $\mathbf{W}_L \in \text{span}(\mathbf{P}_{\leq k})$, by Lemma 20 we have

$$\mathbb{E}_n\left[(f_L(\mathbf{x}; \mathbf{W}_L))^2\right] \lesssim \|f_L(\mathbf{x}; \mathbf{W}_L)\|_{L^2}^2 = \mathcal{R}_2(\mathbf{W}),$$

and similarly

$$\mathbb{E}_n\left[(f_L(\mathbf{x}; \mathbf{W}_L^*))^2\right] \lesssim \mathbb{E}_n\left[(f_k(\mathbf{x})^2\right] + \mathbb{E}_n\left[(f_L(\mathbf{x}; \mathbf{W}_L^*) - f_k(\mathbf{x}))^2\right] \lesssim 1.$$

Altogether

$$\mathbb{E}_n\left[(f_L(\mathbf{x}; \mathbf{\Delta}))^2\right] \lesssim \mathcal{R}_2(\mathbf{W}) + \mathcal{R}_3(\mathbf{W}) \lesssim d^{2(k+1)/3}\varepsilon_{min}^{-4/3}.$$

Plugging this back in,

$$\begin{aligned}\max_{i\in[n]} |\langle\nabla_{\mathbf{W}} f(\mathbf{x}; \mathbf{W}), \mathbf{\Delta}\rangle| &\lesssim n^{1/2}d^{\frac{k+1}{3}}\varepsilon_{min}^{-2/3} + n^{\frac{1}{2}}d \cdot d^{\frac{2k-1}{3}}\varepsilon_{min}^{-1/3} \\ &\lesssim n^{\frac{1}{2}}d^{\frac{2k+2}{3}}\varepsilon_{min}^{-1/3},\end{aligned}$$

since $d^{-k} \ll \varepsilon_{min}$. Therefore

$$\begin{aligned}&\left|\langle\nabla\hat{L}(\mathbf{W}), \mathbf{W} - 2\mathbf{W}_L^* + \mathbf{W}_L\rangle - \langle\nabla\hat{L}^Q(\mathbf{W}), \mathbf{W} - 2\mathbf{W}_L^* + \mathbf{W}_L\rangle\right| \\ &\lesssim d^{3/2}m^{-1/4} \cdot d^{k-\frac{1}{2}}\varepsilon_{min}^{-1/2} \cdot n^{\frac{1}{2}}d^{\frac{2k+2}{3}}\varepsilon_{min}^{-1/3} \\ &= m^{-1/4}n^{\frac{1}{2}}d^{\frac{5(k+1)}{3}}\varepsilon_{min}^{-5/6} \\ &\leq \varepsilon_{min},\end{aligned}$$

since $m \geq n^2 d^{\frac{20(k+1)}{3}} \varepsilon_{min}^{-22/3}$.

Hessian Coupling. By Lemma 24, we have

$$\begin{aligned}
&\left|\mathbb{E}_{\mathbf{S}}\left[\nabla^2 \hat{L}(\mathbf{W})[\mathbf{W}^*\mathbf{S}, \mathbf{W}^*\mathbf{S}]\right] - \mathbb{E}_{\mathbf{S}}\left[\nabla^2 \hat{L}^Q(\mathbf{W})[\mathbf{W}^*\mathbf{S}, \mathbf{W}^*\mathbf{S}]\right]\right| \\
&\leq \mathbb{E}_{\mathbf{S}}\left|\nabla^2 \hat{L}(\mathbf{W})[\mathbf{W}^*\mathbf{S}, \mathbf{W}^*\mathbf{S}] - \nabla^2 \hat{L}^Q(\mathbf{W})[\mathbf{W}^*\mathbf{S}, \mathbf{W}^*\mathbf{S}]\right| \\
&\lesssim d^{3/2} m^{-\frac{1}{4}} \left(\|\mathbf{W}\|_{2,4}^3 + \|\mathbf{W}^*\|_{2,4}^3\right) \left(d\|\mathbf{W}^*\|_{2,4}^2 + \mathbb{E}_{\mathbf{S}} \max_{i\in[n]} |\langle \nabla_{\mathbf{W}} f(\mathbf{x}_i; \mathbf{W}), \mathbf{W}^*\mathbf{S}\rangle|^2\right) \\
&\quad + d^3 \|\mathbf{W}\|_{2,4}^4 \|\mathbf{W}^*\|_{2,\infty}^2.
\end{aligned}$$

By Lemma 27,

$$\begin{aligned}
&\mathbb{E}_{\mathbf{S}} \max_{i\in[n]} |\langle \nabla_{\mathbf{W}} f(\mathbf{x}_i; \mathbf{W}), \mathbf{W}^*\mathbf{S}\rangle|^2 \\
&\lesssim n\mathbb{E}_{\mathbf{S}}\mathbb{E}_n\left[(f_L(\mathbf{x}; \mathbf{W}^*\mathbf{S}))^2\right] + d^2 n m^{-1} \|\mathbf{W}\|_F^2 \|\mathbf{W}^*\|_F^2 \\
&\lesssim n \cdot m^{-1} \|\mathbf{W}^*\|_F^2 + d^2 n \|\mathbf{W}\|_{2,4}^2 \|\mathbf{W}^*\|_{2,4}^2 \\
&\lesssim n d^{\frac{7k+7}{6}} \varepsilon_{min}^{-1/3},
\end{aligned}$$

where we used Lemma 19 to bound $\mathbb{E}_{\mathbf{S}}\mathbb{E}_n\left[(f_L(\mathbf{x}; \mathbf{W}^*\mathbf{S}))^2\right]$, and then plugged in the bound for $\|\mathbf{W}\|_{2,4}$ from Lemma 28 and the bound for $\|\mathbf{W}^*\|_{2,4}$ from Theorem 2.

By Corollary 4, we can bound $\|\mathbf{W}^*\|_{2,\infty} \lesssim m^{-\frac{1}{4}} d^{\frac{k-1}{2}}$. Plugging these two bounds in, along with the bound for $\|\mathbf{W}\|_{2,4}$, we get

$$\begin{aligned}
&\left|\mathbb{E}_{\mathbf{S}}\left[\nabla^2 \hat{L}(\mathbf{W})[\mathbf{W}^*\mathbf{S}, \mathbf{W}^*\mathbf{S}]\right] - \mathbb{E}_{\mathbf{S}}\left[\nabla^2 \hat{L}^Q(\mathbf{W})[\mathbf{W}^*\mathbf{S}, \mathbf{W}^*\mathbf{S}]\right]\right| \\
&\leq d^{3/2} m^{-\frac{1}{4}} d^{k-\frac{1}{2}} \varepsilon_{min}^{-1/2} \cdot n d^{\frac{7k+7}{6}} \varepsilon_{min}^{-1/3} + d^3 \cdot d^{\frac{4k-2}{3}} \varepsilon_{min}^{-2/3} \cdot m^{-1/2} d^{k-1} \\
&\leq m^{-\frac{1}{4}} n d^{\frac{13(k+1)}{6}} \varepsilon_{min}^{-5/6} + m^{-1/2} d^{\frac{(7k+4)}{3}} \varepsilon_{min}^{-2/3} \\
&\leq \varepsilon_{min},
\end{aligned}$$

since we've assumed $m \geq n^4 d^{\frac{26(k+1)}{3}} \varepsilon_{min}^{-22/3}$.

C.4.5 Proof of Lemma 31

Proof. Let us first consider the regularizer $\mathcal{R}_4(\mathbf{W})$. From the proof of [7, Corollary 3], we have

$$\mathbb{E}_{\mathbf{S}} \nabla^2 \mathcal{R}_4(\mathbf{W})[\mathbf{W}^*\mathbf{S}, \mathbf{W}^*\mathbf{S}] - \langle \nabla \mathcal{R}_4(\mathbf{W}), \mathbf{W}\rangle + 2\mathcal{R}_4(\mathbf{W}) - 2\mathcal{R}_4(\mathbf{W}^*) \leq -\mathcal{R}_4(\mathbf{W}) + C\mathcal{R}_4(\mathbf{W}^*)$$

for an absolute constant C. Thus

$$\begin{aligned}
&\mathbb{E}_{\mathbf{S}} \nabla^2 \mathcal{R}_4(\mathbf{W})[\mathbf{W}^*\mathbf{S}, \mathbf{W}^*\mathbf{S}] - \langle \nabla \mathcal{R}_4(\mathbf{W}), \mathbf{W} - 2\mathbf{W}_L^* + \mathbf{W}_L\rangle + 2\mathcal{R}_4(\mathbf{W}) - 2\mathcal{R}_4(\mathbf{W}^*) \\
\leq\ & -\mathcal{R}_4(\mathbf{W}) + C\mathcal{R}_4(\mathbf{W}^*) + 8\|\mathbf{W}\|_{2,4}^4 \sum \|\mathbf{w}_r\|^3 (2\|\{\mathbf{W}_L^*\}_r\| + \|\{\mathbf{W}_L\}_r\|) \\
\leq\ & -\mathcal{R}_4(\mathbf{W}) + C\mathcal{R}_4(\mathbf{W}^*) + 8\|\mathbf{W}\|_{2,4}^7 (2\|\mathbf{W}_L^*\|_{2,4} + \|\mathbf{W}_L\|_{2,4})
\end{aligned}$$

By Lemma 21, we can bound

$$\begin{aligned}
\|\mathbf{W}_L^*\|_{2,4} + \|\mathbf{W}_L\|_{2,4} &\leq m^{-1/4} d^{k/2} \left(\|\mathbf{W}_L^*\|_F + \|\mathbf{W}_L\|_F\right) \\
&\leq m^{-1/4} d^{k/2} d^{\frac{5k-1}{6}} \varepsilon_{min}^{-2/3} \\
&\leq m^{-1/4} d^{\frac{4k}{3} - \frac{1}{6}} \varepsilon_{min}^{-2/3}.
\end{aligned}$$

Plugging this in, along with the bound for $\|\mathbf{W}\|_{2,4}$, yields

$$\begin{aligned}
&\mathbb{E}_{\mathbf{S}} \nabla^2 \mathcal{R}_4(\mathbf{W})[\mathbf{W}^*\mathbf{S}, \mathbf{W}^*\mathbf{S}] - \langle \nabla \mathcal{R}_4(\mathbf{W}), \mathbf{W} - 2\mathbf{W}_L^* + \mathbf{W}_L\rangle + 2\mathcal{R}_4(\mathbf{W}) - 2\mathcal{R}_4(\mathbf{W}^*) \\
&\lesssim \mathcal{R}_4(\mathbf{W}^*) + d^{\frac{7k}{3} - \frac{7}{6}} \varepsilon_{min}^{-7/6} \cdot m^{-1/4} d^{\frac{4k}{3} - \frac{1}{6}} \varepsilon_{min}^{-2/3} \\
&\lesssim \mathcal{R}_4(\mathbf{W}^*) + m^{-1/4} d^{\frac{11k-4}{3}} \varepsilon_{min}^{-11/6}
\end{aligned}$$

and thus

$$\begin{aligned}
&\mathbb{E}_{\mathbf{S}}\nabla^2\lambda_4\mathcal{R}_4(\mathbf{W})[\mathbf{W}^*\mathbf{S},\mathbf{W}^*\mathbf{S}] - \langle\nabla\lambda_4\mathcal{R}_4(\mathbf{W}),\mathbf{W}-2\mathbf{W}_L^*+\mathbf{W}_L\rangle + 2\lambda_4\mathcal{R}_4(\mathbf{W}) - 2\lambda_4\mathcal{R}_4(\mathbf{W}^*)\\
&\lesssim \lambda_4\mathcal{R}_4(\mathbf{W}^*) + \lambda_4 m^{-1/4}d^{\frac{11k-4}{3}}\varepsilon_{min}^{-11/6}\\
&\lesssim \varepsilon_{min} + m^{-1/4}d^{\frac{5k+2}{3}}\varepsilon_{min}^{-5/6}\\
&\lesssim \varepsilon_{min},
\end{aligned}$$

where we used $\lambda_4\mathcal{R}_4(\mathbf{W}^*) \le \varepsilon_{min}$ and $\lambda_4 = d^{-2(k-1)}\varepsilon_{min}$, and then used the assumption $m \ge d^{(20k+8)/3}\varepsilon_{min}^{-22/3}$.

We next deal with the other 3 regularizers. Observe that we can write

$$\mathcal{R}_{tot}(\mathbf{W}) := \lambda_1\mathcal{R}_1(\mathbf{W}) + \lambda_2\mathcal{R}_2(\mathbf{W}) + \lambda_3\mathcal{R}_3(\mathbf{W}) = \text{vec}(\mathbf{W})^T\mathbf{A}\text{vec}(\mathbf{W})$$

for some psd $\mathbf{A} \in \mathbb{R}^{md\times md}$. We get that

$$\begin{aligned}
&\mathbb{E}_{\mathbf{S}}\nabla^2\mathcal{R}_{tot}(\mathbf{W})[\mathbf{W}^*\mathbf{S},\mathbf{W}^*\mathbf{S}] - \langle\nabla\mathcal{R}_{tot}(\mathbf{W}),\mathbf{W}-2\mathbf{W}_L^*+\mathbf{W}_L\rangle + 2\mathcal{R}_{tot}(\mathbf{W}) - 2\mathcal{R}_{tot}(\mathbf{W}^*)\\
\le\ & 2\mathbb{E}_{\mathbf{S}}\mathcal{R}_{tot}(\mathbf{W}^*\mathbf{S}) - 2\mathcal{R}_{tot}(\mathbf{W}) + 4\text{vec}(\mathbf{W})^T\mathbf{A}\text{vec}(\mathbf{W}_L^*) - 2\text{vec}(\mathbf{W})^T\mathbf{A}\text{vec}(\mathbf{W}_L) + 2\mathcal{R}_{tot}(\mathbf{W}) - 2\mathcal{R}_{tot}(\mathbf{W}^*)\\
\le\ & 2\mathbb{E}_{\mathbf{S}}\mathcal{R}_{tot}(\mathbf{W}^*\mathbf{S}) + 4\text{vec}(\mathbf{W})^T\mathbf{A}\text{vec}(\mathbf{W}_L^*) - 2\text{vec}(\mathbf{W})^T\mathbf{A}\text{vec}(\mathbf{W}_L) - 2\mathcal{R}_{tot}(\mathbf{W}_L^*).
\end{aligned}$$

Since $\mathbf{W}_L, \mathbf{W}_L^* \in \text{span}(\mathbf{P}_{\le k})$, we get that

$$\text{vec}(\mathbf{W})^T\mathbf{A}\text{vec}(\mathbf{W}_L) = \lambda_2\text{vec}(\mathbf{W})^T\mathbf{\Sigma}_{\le k}\text{vec}(\mathbf{W}),$$

and

$$\text{vec}(\mathbf{W})^T\mathbf{A}\text{vec}(\mathbf{W}_L^*) = \lambda_2\text{vec}(\mathbf{W})^T\mathbf{\Sigma}_{\le k}\text{vec}(\mathbf{W}_L^*).$$

Therefore

$$\begin{aligned}
4\text{vec}(\mathbf{W})^T\mathbf{A}\text{vec}(\mathbf{W}_L^*) - 2\text{vec}(\mathbf{W})^T\mathbf{A}\text{vec}(\mathbf{W}_L) &\le 2\lambda_2\text{vec}(\mathbf{W}_L^*)^T\mathbf{\Sigma}_{\le k}\text{vec}(\mathbf{W}_L^*)\\
&= 2\lambda_2\mathcal{R}(\mathbf{W}_L^*)\\
&\lesssim \lambda_2\\
&\lesssim \varepsilon_{min}.
\end{aligned}$$

Also,

$$\begin{aligned}
&\mathbb{E}_{\mathbf{S}}\mathcal{R}_{tot}(\mathbf{W}^*\mathbf{S})\\
&= \lambda_1\mathbb{E}_{\mathbf{S}}\mathbb{E}_{\mu}\left[(f_L(\mathbf{x};\mathbf{P}_{>k}\mathbf{W}^*\mathbf{S}))^2\right] + \lambda_2\mathbb{E}_{\mathbf{S}}\mathbb{E}_{\mu}\left[(f_L(\mathbf{x};\mathbf{P}_{\le k}\mathbf{W}^*\mathbf{S}))^2\right] + \lambda_3\mathbb{E}_{\mathbf{S}}\mathbb{E}_{n}\left[(f_L(\mathbf{x};\mathbf{P}_{>k}\mathbf{W}^*\mathbf{S}))^2\right]\\
&\le m^{\frac{1}{2}}d^{-\frac{k-1}{2}}\varepsilon_{min}\left(\mathbb{E}_{\mathbf{S}}\mathbb{E}_{\mu}\left[(f_L(\mathbf{x};\mathbf{W}^*\mathbf{S}))^2\right] + \mathbb{E}_{\mathbf{S}}\mathbb{E}_{n}\left[(f_L(\mathbf{x};\mathbf{W}^*\mathbf{S}))^2\right]\right)\\
&\lesssim m^{\frac{1}{2}}d^{-\frac{k-1}{2}}\varepsilon_{min}\cdot\frac{1}{m}\|\mathbf{W}^*\|_F^2\\
&\lesssim \varepsilon_{min}\cdot d^{-\frac{k-1}{2}}\|\mathbf{W}^*\|_{2,4}^2\\
&\lesssim \varepsilon_{min}.
\end{aligned}$$

Therefore

$$\mathbb{E}_{\mathbf{S}}\nabla^2\mathcal{R}_{tot}(\mathbf{W})[\mathbf{W}^*\mathbf{S},\mathbf{W}^*\mathbf{S}] - \langle\nabla\mathcal{R}_{tot}(\mathbf{W}),\mathbf{W}-2\mathbf{W}_L^*+\mathbf{W}_L\rangle + 2\mathcal{R}_{tot}(\mathbf{W}) - 2\mathcal{R}_{tot}(\mathbf{W}^*) \lesssim \varepsilon_{min}. \tag{68}$$

Summing this with the effect of $\mathcal{R}_4$ on the landscape, we obtain

$$\mathbb{E}_{\mathbf{S}}\left[\nabla^2\mathcal{R}(\mathbf{W})[\mathbf{W}^*\mathbf{S},\mathbf{W}^*\mathbf{S}]\right] - \langle\nabla\mathcal{R}(\mathbf{W}),\mathbf{W}-2\mathbf{W}_L^*+\mathbf{W}_L\rangle + 2\mathcal{R}(\mathbf{W}) - 2\mathcal{R}(\mathbf{W}^*) \lesssim \varepsilon_{min},$$

as desired. □

D Optimization Proofs

D.1 Geometric Properties

In this section we show that the regularized loss L_λ is ℓ-smooth and ρ-Hessian-Lipschitz inside a norm ball.

Lemma 32 (Loss Hessians are Lipschitz)**.**

$$\left\|\nabla^2 \hat{L}(\mathbf{W}_1) - \nabla^2 \hat{L}(\mathbf{W}_2)\right\|_{op} \le d^{3/2}\|\mathbf{W}_1 - \mathbf{W}_2\|_F.$$

Proof. Recall that

$$\begin{aligned}\nabla^2 \hat{L}(\mathbf{W})[\tilde{\mathbf{W}}, \tilde{\mathbf{W}}] = \ & \mathbb{E}_n\left[\ell'(y, f(\mathbf{x}; \mathbf{W}))\frac{1}{\sqrt{m}}\sum_{r=1}^m a_r \sigma''((\mathbf{w}_{0,r} + \mathbf{w}_r)^T\mathbf{x})(\tilde{\mathbf{w}}_r^T\mathbf{x})^2\right] \\ & + \mathbb{E}_n\left[\ell''(y, f(\mathbf{x}; \mathbf{W}))\langle \nabla_{\mathbf{W}} f(\mathbf{x}; \mathbf{W}), \tilde{\mathbf{W}}\rangle^2\right].\end{aligned}$$

Thus

$$\begin{aligned}&\left|\left(\nabla^2 \hat{L}(\mathbf{W}_1) - \nabla^2 \hat{L}(\mathbf{W}_2)\right)[\tilde{\mathbf{W}}, \tilde{\mathbf{W}}]\right| \\ &\le \frac{1}{\sqrt{m}}\sum_{r=1}^m \mathbb{E}_n\left|(\tilde{\mathbf{w}}_r^T\mathbf{x})^2\left(\ell'(y, f(\mathbf{x}; \mathbf{W}_1))\sigma''((\mathbf{w}_{0,r} + \mathbf{w}_{1,r}^T\mathbf{x})) - \ell'(y, f(\mathbf{x}; \mathbf{W}_2))\sigma''((\mathbf{w}_{0,r} + \mathbf{w}_{2,r}^T\mathbf{x}))\right)\right| \\ &+ \mathbb{E}_n\left|\ell''(y, f(\mathbf{x}; \mathbf{W}_1))\langle \nabla_{\mathbf{W}} f(\mathbf{x}; \mathbf{W}_1), \tilde{\mathbf{W}}\rangle^2 - \ell''(y, f(\mathbf{x}; \mathbf{W}_2))\langle \nabla_{\mathbf{W}} f(\mathbf{x}; \mathbf{W}_2), \tilde{\mathbf{W}}\rangle^2\right|\end{aligned}$$

To bound the first term, since ℓ', σ'' are both Lipschitz and can be upper bounded by 1, we get

$$\begin{aligned}&\left|\ell'(y, f(\mathbf{x}; \mathbf{W}_1))\sigma''((\mathbf{w}_{0,r} + \mathbf{w}_{1,r}^T\mathbf{x})) - \ell'(y, f(\mathbf{x}; \mathbf{W}_2))\sigma''((\mathbf{w}_{0,r} + \mathbf{w}_{2,r}^T\mathbf{x}))\right| \\ &\le \left|\ell'(y, f(\mathbf{x}; \mathbf{W}_1)) - \ell'(y, f(\mathbf{x}; \mathbf{W}_2))\right| + \left|\sigma''((\mathbf{w}_{0,r} + \mathbf{w}_{1,r}^T\mathbf{x})) - \sigma''((\mathbf{w}_{0,r} + \mathbf{w}_{2,r}^T\mathbf{x}))\right| \\ &\le \left|f(\mathbf{x}; \mathbf{W}_1) - f(\mathbf{x}; \mathbf{W}_2)\right| + \left|(\mathbf{w}_{1,r} - \mathbf{w}_{2,r})^T\mathbf{x}\right| \\ &\le \left|(\mathbf{w}_{1,r} - \mathbf{w}_{2,r})^T\mathbf{x}\right| + \frac{1}{\sqrt{m}}\sum_{s=1}^m \left|\sigma((\mathbf{w}_{0,s} + \mathbf{w}_{1,s}^T\mathbf{x})) - \sigma((\mathbf{w}_{0,s} + \mathbf{w}_{2,s}^T\mathbf{x}))\right| \\ &\le \left|(\mathbf{w}_{1,r} - \mathbf{w}_{2,r})^T\mathbf{x}\right| + \frac{1}{\sqrt{m}}\sum_{s=1}^m \left|(\mathbf{w}_{1,s} - \mathbf{w}_{2,s})^T\mathbf{x}\right|\end{aligned}$$

Therefore

$$\begin{aligned}&\frac{1}{\sqrt{m}}\sum_{r=1}^m \mathbb{E}_n\left|(\tilde{\mathbf{w}}_r^T\mathbf{x})^2\left(\ell'(y, f(\mathbf{x}; \mathbf{W}_1))\sigma''((\mathbf{w}_{0,r} + \mathbf{w}_{1,r}^T\mathbf{x})) - \ell'(y, f(\mathbf{x}; \mathbf{W}_2))\sigma''((\mathbf{w}_{0,r} + \mathbf{w}_{2,r}^T\mathbf{x}))\right)\right| \\ &\le \frac{1}{\sqrt{m}}\mathbb{E}_n\left|(\tilde{\mathbf{w}}_r^T\mathbf{x})^2(\mathbf{w}_{1,r} - \mathbf{w}_{2,r})^T\mathbf{x}\right| + \frac{1}{m}\sum_{r,s=1}^m \mathbb{E}_n\left|(\tilde{\mathbf{w}}_r^T\mathbf{x})^2(\mathbf{w}_{1,s} - \mathbf{w}_{2,s})^T\mathbf{x}\right| \\ &\lesssim \frac{d^{3/2}}{\sqrt{m}}\sum_{r=1}^m \|\tilde{\mathbf{w}}_r\|^2\|\mathbf{w}_{1,r} - \mathbf{w}_{2,r}\| + \frac{d^{3/2}}{m}\|\tilde{\mathbf{W}}\|_F^2\sum_{s=1}^m \|\mathbf{w}_{1,s} - \mathbf{w}_{2,s}\| \\ &\lesssim \frac{d^{3/2}}{\sqrt{m}}\|\tilde{\mathbf{W}}\|_F^2\|\mathbf{W}_1 - \mathbf{W}_2\|_F.\end{aligned}$$

To bound the second term, we have

$$\begin{aligned}&\left|\ell''(y, f(\mathbf{x}; \mathbf{W}_1))\langle \nabla_{\mathbf{W}} f(\mathbf{x}; \mathbf{W}_1), \tilde{\mathbf{W}}\rangle^2 - \ell''(y, f(\mathbf{x}; \mathbf{W}_2))\langle \nabla_{\mathbf{W}} f(\mathbf{x}; \mathbf{W}_2), \tilde{\mathbf{W}}\rangle^2\right| \\ &\le \left|\langle \nabla_{\mathbf{W}} f(\mathbf{x}; \mathbf{W}_1), \tilde{\mathbf{W}}\rangle^2 - \langle \nabla_{\mathbf{W}} f(\mathbf{x}; \mathbf{W}_2), \tilde{\mathbf{W}}\rangle^2\right| + \langle \nabla_{\mathbf{W}} f(\mathbf{x}; \mathbf{W}_2), \tilde{\mathbf{W}}\rangle^2\left|\ell''(y, f(\mathbf{x}; \mathbf{W}_1)) - \ell''(y, f(\mathbf{x}; \mathbf{W}_2))\right| \\ &\le \|\tilde{\mathbf{W}}\|_F^2\|\nabla_{\mathbf{W}} f(\mathbf{x}; \mathbf{W}_1) - \nabla_{\mathbf{W}} f(\mathbf{x}; \mathbf{W}_2)\|\|\nabla_{\mathbf{W}} f(\mathbf{x}; \mathbf{W}_1) + \nabla_{\mathbf{W}} f(\mathbf{x}; \mathbf{W}_2)\| \\ &\quad + \|\tilde{\mathbf{W}}\|_F^2\|\nabla_{\mathbf{W}} f(\mathbf{x}; \mathbf{W}_2)\|^2|f(\mathbf{x}; \mathbf{W}_1) - f(\mathbf{x}; \mathbf{W}_2)|.\end{aligned}$$

Since σ' is bounded by 1, we can bound $\|\nabla_{\mathbf{W}} f(\mathbf{x}; \mathbf{W}_2)\| \le \sqrt{d}$. Next, we have

$$\begin{aligned}\|\nabla_{\mathbf{W}} f(\mathbf{x}; \mathbf{W}_1) - \nabla_{\mathbf{W}} f(\mathbf{x}; \mathbf{W}_2)\|_F^2 &\le \frac{1}{m}\|\sigma'((\mathbf{w}_{0,r}^+ \mathbf{w}_{1,r})^T \mathbf{x})\mathbf{x} - \sigma'((\mathbf{w}_{0,r}^+ \mathbf{w}_{2,r})^T \mathbf{x})\mathbf{x}\|^2 \\ &\le \frac{d}{m}\sum_{r=1}^m \left|(\mathbf{w}_{1,r} - \mathbf{w}_{2,r})^T \mathbf{x}\right|^2 \\ &\le \frac{d^2}{m}\|\mathbf{W}_1 - \mathbf{W}_2\|_F^2.\end{aligned}$$

Finally,

$$\begin{aligned}|f(\mathbf{x}; \mathbf{W}_1) - f(\mathbf{x}; \mathbf{W}_2)| &\le \frac{1}{\sqrt{m}}\sum_{r=1}^m \left|(\mathbf{w}_{1,r} - \mathbf{w}_{2,r})^T \mathbf{x}\right| \\ &\le \sqrt{d}\|\mathbf{W}_1 - \mathbf{W}_2\|_F.\end{aligned}$$

Altogether,

$$\begin{aligned}&\left|\ell''(y, f(\mathbf{x}; \mathbf{W}_1))\langle \nabla_{\mathbf{W}} f(\mathbf{x}; \mathbf{W}_1), \tilde{\mathbf{W}}\rangle^2 - \ell''(y, f(\mathbf{x}; \mathbf{W}_2))\langle \nabla_{\mathbf{W}} f(\mathbf{x}; \mathbf{W}_2), \tilde{\mathbf{W}}\rangle^2\right| \\ &\le \|\tilde{\mathbf{W}}\|_F^2 d^{3/2}\|\mathbf{W}_1 - \mathbf{W}_2\|_F.\end{aligned}$$

Therefore

$$\left|\left(\nabla^2 \hat{L}(\mathbf{W}_1) - \nabla^2 \hat{L}(\mathbf{W}_2)\right)[\tilde{\mathbf{W}}, \tilde{\mathbf{W}}]\right| \le \|\tilde{\mathbf{W}}\|_F^2 d^{3/2}\|\mathbf{W}_1 - \mathbf{W}_2\|_F,$$

so

$$\left\|\nabla^2 \hat{L}(\mathbf{W}_1) - \nabla^2 \hat{L}(\mathbf{W}_2)\right\|_{op} \le d^{3/2}\|\mathbf{W}_1 - \mathbf{W}_2\|_F.$$

□

Lemma 33 (Regularized loss is Hessian-Lipschitz)**.** *The regularized loss L_λ is $O(\lambda_4 \Gamma^5)$-Hessian-Lipschitz inside the region $\{\mathbf{W} \mid \|\mathbf{W}\|_F \le \Gamma\}$*

Proof. For $i = 1, 2, 3$, the regularizer $\mathcal{R}_i$ is a convex quadratic, so $\nabla^2 \mathcal{R}_i(\mathbf{W}_1) = \nabla^2 \mathcal{R}_i(\mathbf{W}_2)$. As for the regularizer $\mathcal{R}_4$, we have

$$\nabla^2 \mathcal{R}_4(\mathbf{W})[\tilde{\mathbf{W}}, \tilde{\mathbf{W}}] = 32\left(\sum_{r=1}^m \|\mathbf{w}_r\|^2 \mathbf{w}_r^T \tilde{\mathbf{w}}_r\right)^2 + 8\|\mathbf{W}\|_{2,4}^4 \left(\sum_{r=1}^m 2(\mathbf{w}_r^T \tilde{\mathbf{w}})^2 + \|\mathbf{w}_r\|^2 \|\tilde{\mathbf{w}}_r\|^2\right).$$

Therefore, for $\|\tilde{\mathbf{W}}\|_F = 1$,

$$\begin{aligned}&\left|\nabla^2 \mathcal{R}_4(\mathbf{W}_1)[\tilde{\mathbf{W}}, \tilde{\mathbf{W}}] - \nabla^2 \mathcal{R}_4(\mathbf{W}_2)[\tilde{\mathbf{W}}, \tilde{\mathbf{W}}]\right| \\ &\le 32\left(\sum_{r=1}^m (\|\mathbf{w}_{1,r}\|^2 \mathbf{w}_{1,r}^T + \|\mathbf{w}_{2,r}\|^2 \mathbf{w}_{2,r}^T)\tilde{\mathbf{w}}_r\right)\left(\sum_{r=1}^m (\|\mathbf{w}_{1,r}\|^2 \mathbf{w}_{1,r}^T - \|\mathbf{w}_{2,r}\|^2 \mathbf{w}_{2,r}^T)\tilde{\mathbf{w}}_r\right) \\ &\quad + 8\sum_{r=1}^m 2\|\mathbf{W}_1\|_{2,4}^4 (\mathbf{w}_{1,r}^T \tilde{\mathbf{w}}_r)^2 - 2\|\mathbf{W}_2\|_{2,4}^4 (\mathbf{w}_{2,r}^T \tilde{\mathbf{w}}_r)^2 + (\|\mathbf{W}_1\|_{2,4}^4 \|\mathbf{w}_{1,r}\|^2 - \|\mathbf{W}_2\|_{2,4}^4 \|\mathbf{w}_{2,r}\|^2)\|\tilde{\mathbf{w}}_r\|^2.\end{aligned}$$

We can bound the first term using

$$\begin{aligned}\left|\sum_{r=1}^m (\|\mathbf{w}_{1,r}\|^2 \mathbf{w}_{1,r}^T + \|\mathbf{w}_{2,r}\|^2 \mathbf{w}_{2,r}^T)\tilde{\mathbf{w}}_r\right| &\le \sum_{r=1}^m (\|\mathbf{w}_{1,r}\|^3 + \|\mathbf{w}_{2,r}\|^3)\|\tilde{\mathbf{w}}_r\| \\ &\le \left(\|\mathbf{W}_1\|_{2,6}^3 + \|\mathbf{W}_2\|_{2,6}^3\right)\|\tilde{\mathbf{W}}\|_F \\ &= \|\mathbf{W}_1\|_{2,6}^3 + \|\mathbf{W}_2\|_{2,6}^3.\end{aligned}$$

and

$$\begin{aligned}\left|\sum_{r=1}^{m}(\|\mathbf{w}_{1,r}\|^2\mathbf{w}_{1,r}^T-\|\mathbf{w}_{2,r}\|^2\mathbf{w}_{2,r}^T)\tilde{\mathbf{w}}_r\right| &\le \sum_{r=1}^{m}\left|\|\mathbf{w}_{1,r}\|^2\mathbf{w}_{1,r}-\|\mathbf{w}_{2,r}\|^2\mathbf{w}_{2,r}\right|\|\tilde{\mathbf{w}}_r\|\\ &\le \sum_{r=1}^{m}\left|\|\mathbf{w}_{1,r}\|^2\mathbf{w}_{1,r}-\|\mathbf{w}_{2,r}\|^2\mathbf{w}_{2,r}\right|\\ &\lesssim \sum_{r=1}^{m}\|\mathbf{w}_{1,r}-\mathbf{w}_{2,r}\|(\|\mathbf{w}_{1,r}\|^2+\|\mathbf{w}_{2,r}\|^2)\\ &\le \|\mathbf{W}_1-\mathbf{W}_2\|_F(\|\mathbf{W}_1\|_{2,4}^2+\|\mathbf{W}_2\|_{2,4}^2).\end{aligned}$$

For the second term, we bound

$$\begin{aligned}&\left|\sum_{r=1}^{m}\|\mathbf{W}_1\|_{2,4}^4(\mathbf{w}_{1,r}^T\tilde{\mathbf{w}}_r)^2-\|\mathbf{W}_2\|_{2,4}^4(\mathbf{w}_{2,r}^T\tilde{\mathbf{w}}_r)^2\right|\\ &\le \sum_{r=1}^{m}\left|\|\mathbf{W}_1\|_{2,4}^4\mathbf{w}_{1,r}\mathbf{w}_{1,r}^T-\|\mathbf{W}_2\|_{2,4}^4\mathbf{w}_{2,r}\mathbf{w}_{2,r}^T\right|_{op}\\ &\le \sum_{r=1}^{m}\left(\|\mathbf{W}_1\|_{2,4}^2\|\mathbf{w}_{1,r}\|+\|\mathbf{W}_2\|_{2,4}^2\|\mathbf{w}_{2,r}\|\right)\left(\left|\|\mathbf{W}_1\|_{2,4}^2\mathbf{w}_{1,r}-\|\mathbf{W}_2\|_{2,4}^2\mathbf{w}_{2,r}\right|\right)\\ &\lesssim \sum_{r=1}^{m}\left(\|\mathbf{W}_1\|_{2,4}^2\|\mathbf{w}_{1,r}\|+\|\mathbf{W}_2\|_{2,4}^2\|\mathbf{w}_{2,r}\|\right)\left(\|\mathbf{W}_1\|_{2,4}^2+\|\mathbf{W}_2\|_{2,4}^2\right)\|\mathbf{w}_{1,r}-\mathbf{w}_{2,r}\|\\ &\lesssim \left(\|\mathbf{W}_1\|_{2,4}^2+\|\mathbf{W}_2\|_{2,4}^2\right)\left(\|\mathbf{W}_1\|_{2,4}^2\|\mathbf{W}_1\|_F+\|\mathbf{W}_2\|_{2,4}^2\|\mathbf{W}_2\|_F\right)\|\mathbf{W}_1-\mathbf{W}_2\|_F.\end{aligned}$$

Finally, we bound

$$\begin{aligned}&\sum_{r=1}^{m}\left|\|\mathbf{W}_1\|_{2,4}^4\|\mathbf{w}_{1,r}\|^2-\|\mathbf{W}_2\|_{2,4}^4\|\mathbf{w}_{2,r}\|^2\right|\|\tilde{\mathbf{w}}_r\|^2\\ &\le \sum_{r=1}^{m}\left|\|\mathbf{W}_1\|_{2,4}^4\|\mathbf{w}_{1,r}\|^2-\|\mathbf{W}_2\|_{2,4}^4\|\mathbf{w}_{2,r}\|^2\right|\\ &\le \sum_{r=1}^{m}\left|\|\mathbf{W}_1\|_{2,4}^4-\|\mathbf{W}_2\|_{2,4}^4\right|\|\mathbf{w}_{1,r}\|^2+\|\mathbf{W}_2\|_{2,4}^4\left|\|\mathbf{w}_{1,r}\|^2-\|\mathbf{w}_{2,r}\|^2\right|\\ &\le \|\mathbf{W}_1\|_F^2\left|\|\mathbf{W}_1\|_{2,4}^4-\|\mathbf{W}_2\|_{2,4}^4\right|+\|\mathbf{W}_2\|_{2,4}^4\sum_{r=1}^{m}(\|\mathbf{w}_{1,r}\|+\|\mathbf{w}_{2,r}\|)\|\mathbf{w}_{1,r}-\mathbf{w}_{2,r}\|\\ &\lesssim \|\mathbf{W}_1\|_F^2\|\mathbf{W}_1-\mathbf{W}_2\|_{2,4}(\|\mathbf{W}_1\|_{2,4}^3+\|\mathbf{W}_2\|_{2,4}^3)+\|\mathbf{W}_2\|_{2,4}^4(\|\mathbf{W}_1\|_F+\|\mathbf{W}_2\|_F)\|\mathbf{W}_1-\mathbf{W}_2\|_F\\ &\le \left(\|\mathbf{W}_1\|_F^2(\|\mathbf{W}_1\|_{2,4}^3+\|\mathbf{W}_2\|_{2,4}^3)+\|\mathbf{W}_2\|_{2,4}^4(\|\mathbf{W}_1\|_F+\|\mathbf{W}_2\|_F)\right)\|\mathbf{W}_1-\mathbf{W}_2\|_F.\end{aligned}$$

Altogether, when $\|\mathbf{W}_1\|_F, \|\mathbf{W}_2\|_F \le \Gamma$, and using $\|\mathbf{W}\|_{2,2k} \le \|\mathbf{W}\|_F$ for $k \ge 1$, we get that

$$\left|\nabla^2\mathcal{R}_4(\mathbf{W}_1)[\tilde{\mathbf{W}},\tilde{\mathbf{W}}]-\nabla^2\mathcal{R}_4(\mathbf{W}_2)[\tilde{\mathbf{W}},\tilde{\mathbf{W}}]\right| \lesssim \Gamma^5\|\mathbf{W}_1-\mathbf{W}_F\|.$$

Therefore L_λ is $O(\Gamma^5)$-Hessian-Lipschitz. □

Lemma 34 (Regularized loss is smooth)**.** *The regularized loss L_λ is $O(\lambda_4\Gamma^6+m^{1/2})$-smooth.*

Proof. Recall that

$$\begin{aligned}\nabla^2\hat{L}(\mathbf{W})[\tilde{\mathbf{W}},\tilde{\mathbf{W}}] = \ &\mathbb{E}_n\left[\ell'(y,f(\mathbf{x};\mathbf{W}))\frac{1}{\sqrt{m}}\sum_{r=1}^{m}a_r\sigma''((\mathbf{w}_{0,r}+\mathbf{w}_r)^T\mathbf{x})(\tilde{\mathbf{w}}_r^T\mathbf{x})^2\right]\\ &+\mathbb{E}_n\left[\ell''(y,f(\mathbf{x};\mathbf{W}))\langle\nabla_{\mathbf{W}}f(\mathbf{x};\mathbf{W}),\tilde{\mathbf{W}}\rangle^2\right],\end{aligned}$$

and thus, for $\|\tilde{\mathbf{W}}\|_F = 1$,

$$\begin{aligned}\left|\nabla^2 \hat{L}(\mathbf{W})[\tilde{\mathbf{W}}, \tilde{\mathbf{W}}]\right| &\le \frac{1}{\sqrt{m}} \sum_{r=1}^{m} \mathbb{E}_n[(\tilde{\mathbf{w}}_r^T \mathbf{x})^2] + \mathbb{E}_n \|\nabla_{\mathbf{W}} f(\mathbf{x}; \mathbf{W})\|^2 \|\tilde{\mathbf{W}}\|^2 \\ &\lesssim \frac{d}{\sqrt{m}} \|\tilde{\mathbf{W}}\|_F^2 + d\|\tilde{\mathbf{W}}\|^2 \\ &\lesssim d.\end{aligned}$$

Therefore $\hat{L}$ is $O(d)$-smooth.

Next, consider the regularizers. $\mathcal{R}_1, \mathcal{R}_2, \mathcal{R}_3$ are all convex quadratics, and since $\|\varphi(\mathbf{x})\| \le \sqrt{d}$, we can upper bound the smoothness of $\mathcal{R}_1, \mathcal{R}_2, \mathcal{R}_3$ by d. Therefore the smoothness of $\lambda_1\mathcal{R}_1 + \lambda_2\mathcal{R}_2 + \lambda_3\mathcal{R}_3$ is at most $m^{1/2} d^{-\frac{k-1}{2}} \varepsilon_{min} \cdot d \lesssim m^{1/2}$.

Finally, we consider $\mathcal{R}_4$. For $\|\tilde{\mathbf{W}}\|_F = 1$, we can bound

$$\begin{aligned}\nabla^2 \mathcal{R}_4(\mathbf{W})[\tilde{\mathbf{W}}, \tilde{\mathbf{W}}] &= 32 \left(\sum_{r=1}^{m} \|\mathbf{w}_r\|^2 \mathbf{w}_r^T \tilde{\mathbf{w}}_r \right)^2 + 8\|\mathbf{W}\|_{2,4}^4 \left(\sum_{r=1}^{m} 2(\mathbf{w}_r^T \tilde{\mathbf{w}})^2 + \|\mathbf{w}_r\|^2 \|\tilde{\mathbf{w}}_r\|^2 \right) \\ &\le 32 \left(\sum_{r=1}^{m} \|\mathbf{w}_r\|^3 \|\tilde{\mathbf{w}}_r\| \right)^2 + 24\|\mathbf{W}\|_{2,4}^4 \sum_{r=1}^{m} \|\mathbf{w}_r\|^2 \|\tilde{\mathbf{w}}_r\|^2 \\ &\le 32\|\mathbf{W}\|_{2,6}^6 \|\tilde{\mathbf{W}}\|_F^2 + 24\|\mathbf{W}\|_{2,4}^4 \|\mathbf{W}\|_F^2 \\ &\le 32\|\mathbf{W}\|_{2,6}^6 + 24\|\mathbf{W}\|_{2,4}^4 \|\mathbf{W}\|_F^2 \\ &\le 56\|\mathbf{W}\|_F^6.\end{aligned}$$

Therefore L_λ is $O(\lambda_4 \Gamma^6 + m^{1/2})$-smooth. □

D.2 Proof of Theorem 3

We prove the following formal version of Theorem 3.

Theorem 5. *For ν, γ, choose $\eta = \frac{c}{\lambda_4 m^3}$ for sufficiently small constant c and define $\tilde{\varepsilon} = \min(\nu, \gamma^2 m^{-5/2}), \sigma = \tilde{\Theta}(\tilde{\varepsilon})$. Then with probability $1 - d^{-8}$, perturbed gradient descent reaches a (ν, γ)-SOSP within $T = \tilde{O}(m^3 \tilde{\varepsilon}^{-2})$ timesteps.*

Proof. We follow the same strategy as [30, Theorem 8]. We first show that perturbed gradient descent stays in a bounded region. Then, we can use the smoothness and Hessian-Lipschitz parameters in this region for the purpose of a convergence result. Throughout, we condition on the high probability event where the construction in Theorem 2 holds.

Let $\Gamma = m^{1/2}$. I claim that perturbed gradient descent stays in the region $\{\|\mathbf{W}\| \mid \|\mathbf{W}\| \le \Gamma\}$. We prove the claim by induction. $\|\mathbf{W}^0\| = 0$ so clearly the base case holds.

Assume that $\|\mathbf{W}^t\| \le \Gamma$. The gradient descent update on $\mathbf{W}$ is

$$\mathbf{W}^{t+1} \leftarrow \mathbf{W}^t - \eta \nabla \hat{L}(\mathbf{W}^t) - \eta \nabla \left(\lambda_1 \mathcal{R}_1(\mathbf{W}^t) + \lambda_2 \mathcal{R}_2(\mathbf{W}^t) + \lambda_3 \mathcal{R}_3(\mathbf{W}^t) \right) - \eta \lambda_4 \nabla \mathcal{R}_4(\mathbf{W}^t).$$

Observe that the learning rate is $\eta = \frac{c}{\lambda_4 m^3} = \frac{c}{\lambda_4 \Gamma^6}$ for a sufficiently small constant c. We can bound $\|\nabla \hat{L}(\mathbf{W}^t)\| \le \sqrt{d}$ and, as in the proof of the previous lemma,

$$\left\| \nabla \left(\lambda_1 \mathcal{R}_1(\mathbf{W}^t) + \lambda_2 \mathcal{R}_2(\mathbf{W}^t) + \lambda_3 \mathcal{R}_3(\mathbf{W}^t) \right) \right\| \le m^{1/2} \|\mathbf{W}^t\|_F.$$

Therefore

$$\begin{aligned}\left\| \eta \nabla \hat{L}(\mathbf{W}^t) + \eta \nabla \left(\lambda_1 \mathcal{R}_1(\mathbf{W}^t) + \lambda_2 \mathcal{R}_2(\mathbf{W}^t) + \lambda_3 \mathcal{R}_3(\mathbf{W}^t) \right) \right\| &\lesssim \lambda_4^{-1} \Gamma^{-6} \sqrt{d} + \lambda_4^{-1} \Gamma^{-5} m^{1/2} \\ &\lesssim \lambda_4^{-1} \Gamma^{-5} m^{1/2}\end{aligned}$$

Finally, since the $\mathbf{w}_r$ component of $\nabla \mathcal{R}_4(\mathbf{W})$ is $8\|\mathbf{W}\|_{2,4}^4 \|\mathbf{w}_r\|^2 \mathbf{w}_r$, we have

$$\begin{aligned}\{\mathbf{W}^t - \eta\lambda_4 \nabla \mathcal{R}_4(\mathbf{W}^t)\}_r &= \mathbf{w}_r^t \left(1 - 8\lambda_4 \eta \|\mathbf{W}^t\|_{2,4}^4 \|\mathbf{w}_r^t\|^2\right) \\ &= \mathbf{w}_r^t \left(1 - 8c\Gamma^{-6} \|\mathbf{W}^t\|_{2,4}^4 \|\mathbf{w}_r^t\|^2\right)\end{aligned}$$

Note that

$$8c\Gamma^{-6} \|\mathbf{W}^t\|_{2,4}^4 \|\mathbf{w}_r^t\|^2 \leq 8c \leq 1$$

for c small enough. We then have

$$\begin{aligned}\|\mathbf{W}^t - \eta\lambda_4 \nabla \mathcal{R}_4(\mathbf{W}^t)\|_F^2 &= \sum_r \|\mathbf{w}_r^t\|^2 \left(1 - 8c\Gamma^{-6} \|\mathbf{W}^t\|_{2,4}^4 \|\mathbf{w}_r^t\|^2\right)^2 \\ &\leq \sum_r \|\mathbf{w}_r^t\|^2 \left(1 - 8c\Gamma^{-6} \|\mathbf{W}^t\|_{2,4}^4 \|\mathbf{w}_r^t\|^2\right) \\ &= \|\mathbf{W}^t\|_F^2 - 8c\Gamma^{-6} \|\mathbf{W}^t\|_{2,4}^8 \\ &\leq \|\mathbf{W}^t\|_F^2 - 8c\Gamma^{-6} m^{-2} \|\mathbf{W}^t\|_F^8.\end{aligned}$$

We split into two cases. If $\|\mathbf{W}^t\|_F \leq \Gamma/2$, then

$$\|\mathbf{W}^t - \eta\lambda_4 \nabla \mathcal{R}_4(\mathbf{W}^t)\|_F \leq \Gamma/2$$

and

$$\begin{aligned}\left\|\eta \nabla \hat{L}(\mathbf{W}^t) + \eta \nabla \left(\lambda_1 \mathcal{R}_1(\mathbf{W}^t) + \lambda_2 \mathcal{R}_2(\mathbf{W}^t) + \lambda_3 \mathcal{R}_3(\mathbf{W}^t)\right)\right\| &\lesssim d^{2(k-1)} \varepsilon_{min}^{-1} \Gamma^{-5} m^{1/2} \\ &\leq \Gamma/4,\end{aligned}$$

so $\|\mathbf{W}^{t+1}\|_F \leq 3\Gamma/4$.

Otherwise, $\Gamma \geq \|\mathbf{W}^t\|_F \geq \Gamma/2$, so

$$\begin{aligned}\|\mathbf{W}^t - \eta\lambda_4 \nabla \mathcal{R}_4(\mathbf{W}^t)\|_F &\leq \sqrt{\Gamma^2 - \frac{c}{32}\Gamma^2 m^{-2}}. \\ &\leq \Gamma(1 - \frac{c}{64} m^{-2})\end{aligned}$$

Then

$$\begin{aligned}\|\mathbf{W}^{t+1}\|_F &\leq \Gamma(1 - \frac{c}{64} m^{-2}) + d^{2(k-1)} \varepsilon_{min}^{-1} \Gamma^{-5} m^{1/2} \\ &\leq \Gamma\left(1 - \frac{c}{128} m^{-2}\right),\end{aligned}$$

since $\Gamma^6 = m^3 \gtrsim m^{5/2} d^{2(k-1)} \varepsilon_{min}^{-1}$.

Finally, the perturbation moves at most $\eta\|\mathbf{\Xi}^t\|_F$, and since $\mathbb{E}\|\mathbf{\Xi}^t\|_F = \sigma$, with probability $1 - d^{-9}$ each of the T perturbations is bounded by $\Gamma m^{-3} \eta^{-1} \gg \sigma$ Therefore even after the perturbation we have $\|\mathbf{W}^{t+1}\|_F \leq \Gamma$, completing the induction step.

Lemmas 33, 34 tell us L_λ is $O(\lambda_4 m^{5/2})$ Hessian Lipschitz and $O(\lambda_4 m^3)$ smooth throughout the entire gradient descent trajectory.

Our goal is to converge to a (ν, γ)-SOSP; this is equivalent to converging to a $\tilde{\varepsilon}$-SOSP as defined in [30, 31], with $\tilde{\varepsilon} := \min(\nu, \frac{\gamma^2}{\lambda_4 m^{5/2}}) \geq \min(\nu, \gamma^2 m^{-5/2})$. Therefore by [30, 31], with probability $1 - d^{-9}$ perturbed gradient descent on the regularized loss with learning rate $\eta = \frac{c}{\lambda_4 m^3}$ and perturbation radius $\sigma = \tilde{\Theta}(\tilde{\varepsilon})$ will encounter an $\tilde{\varepsilon}$-SOSP in at most $T = \tilde{O}\left(\lambda_4 m^3 \tilde{\varepsilon}^{-2}\right) \leq \tilde{O}\left(m^3 \tilde{\varepsilon}^{-2}\right)$ timesteps. Union bounding over the high probability events, this occurs with probability $1 - 3d^{-9} \geq 1 - d^{-8}$. $\square$

E Generalization Proofs

E.1 Proof of Theorem 4

Recall the definition of the empirical Rademacher complexity:

Definition 4. *Let $\mathcal{F}$ be a class of functions from $\mathbb{R}^d$ to $\mathbb{R}$. Given a dataset $\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, the* empirical Rademacher complexity *of $\mathcal{F}$ is defined as*

$$\mathcal{R}_{\mathcal{D}}(\mathcal{F}) := \mathbb{E}_{\sigma \in \{\pm 1\}^n} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(\mathbf{x}_i) \right]. \tag{69}$$

We next show that the Rademacher complexities of the linear term and the quadratic term can be bounded.

Lemma 35 (Rademacher complexity of linear term)**.** *Let $\mathcal{W} \subset \mathbb{R}^{m \times d}$, and define the function class $\mathcal{F}_k^L(\mathcal{W}) := \{\mathbf{x} \mapsto f^L(\mathbf{x}; \mathbf{P}_{\le k}\mathbf{W}) : \mathbf{W} \in \mathcal{W}\}$. Then, with probability $1 - d^{-9}$ over the draw of $\mathcal{D}$,*

$$\mathcal{R}_{\mathcal{D}}(\mathcal{F}_k^L(\mathcal{W})) \lesssim \sqrt{\frac{d^k}{n}} \cdot \sup_{\mathbf{W} \in \mathcal{W}} \|f_L(\mathbf{x}; \mathbf{P}_{\le k}\mathbf{W})\|_{L^2}^2. \tag{70}$$

Lemma 36 (Rademacher complexity of quadratic term)**.** *Let $\mathcal{W} \subset \mathbb{R}^{m \times d}$, and define the function class $\mathcal{F}^Q(\mathcal{W}) := \{\mathbf{x} \mapsto f^Q(\mathbf{x}; \mathbf{W}) : \mathbf{W} \in \mathcal{W}\}$. Then, with probability $1 - d^{-9}$ over the draw of $\mathcal{D}$,*

$$\mathcal{R}_{\mathcal{D}}(\mathcal{F}^Q(\mathcal{W})) \lesssim \sqrt{\frac{d}{mn}} \cdot \sup_{\mathbf{W} \in \mathcal{W}} \|\mathbf{W}\|_F^2. \tag{71}$$

Lemma 35 is presented in Appendix E.2; Lemma 36 follows directly from [7, Lemma 5, Theorem 6].

Equipped with these Rademacher complexity lemmas, we can now prove the main generalization result.

Proof of Theorem 4. By Lipschitzness of the loss and Lemma 25, we can bound

$$\begin{aligned}
L(\mathbf{W}) &= \mathbb{E}_\mu[\ell(y, f(\mathbf{x}; \mathbf{W}))] \\
&\le \mathbb{E}_\mu[\ell(y, f_Q(\mathbf{x}; \mathbf{W}) + f_L(\mathbf{x}; \mathbf{W}))] + \mathbb{E}_\mu|f(\mathbf{x}; \mathbf{W}) - f_Q(\mathbf{x}; \mathbf{W}) - f_L(\mathbf{x}; \mathbf{W})| \\
&\le \mathbb{E}_\mu[\ell(y, f_Q(\mathbf{x}; \mathbf{W}) + f_L(\mathbf{x}; \mathbf{W}))] + m^{-\frac{1}{2}} \sum_{r=1}^{m} \mathbb{E}_\mu|\mathbf{w}_r^T \mathbf{x}|^3 \\
&\le \mathbb{E}_\mu[\ell(y, f_Q(\mathbf{x}; \mathbf{W}) + f_L(\mathbf{x}; \mathbf{W}))] + Cm^{-\frac{1}{4}} \|\mathbf{W}\|_{2,4}^3.
\end{aligned}$$

Again by Lipschitzness, we can bound

$$\begin{aligned}
\mathbb{E}_\mu[\ell(y, f_Q(\mathbf{x}; \mathbf{W}) + f_L(\mathbf{x}; \mathbf{W}))] &\le \mathbb{E}_\mu[\ell(y, f_Q(\mathbf{x}; \mathbf{W}) + f_L(\mathbf{x}; \mathbf{P}_{\le k}\mathbf{W}))] + \mathbb{E}_\mu|f_L(x; \mathbf{P}_{>k}\mathbf{W})| \\
&\le \mathbb{E}_\mu[\ell(y, f_Q(\mathbf{x}; \mathbf{W}) + f_L(\mathbf{x}; \mathbf{P}_{\le k}\mathbf{W}))] + \|f_L(\mathbf{x}; \mathbf{P}_{>k}\mathbf{W})\|_{L^2}.
\end{aligned}$$

Similarly, we can lower bound

$$\hat{L}(\mathbf{W}) \ge \mathbb{E}_n[\ell(y, f_Q(\mathbf{x}; \hat{\mathbf{W}}) + f_L(\mathbf{x}; \mathbf{P}_{\le k}\hat{\mathbf{W}}))] - Cm^{-\frac{1}{4}} \|\mathbf{W}\|_{2,4}^3 - \left(\mathbb{E}_n[(f_L(\mathbf{x}; \mathbf{P}_{>k}))^2]\right)^{\frac{1}{2}}.$$

Since $L_\lambda(\hat{\mathbf{W}}) \le C\varepsilon_{min}$, the value of each regularizer satisfies $\mathcal{R}_i(\hat{\mathbf{W}}) \le C\lambda_i^{-1}\varepsilon_{min}$. By our choice of $(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$, we have

$$\begin{aligned}
\mathcal{R}_1(\hat{\mathbf{W}}) &= \|f_L(\cdot; \mathbf{P}_{>k}\mathbf{W})\|_{L^2}^2 \lesssim m^{-\frac{1}{2}} d^{\frac{k-1}{2}}, \\
\mathcal{R}_2(\hat{\mathbf{W}}) &= \|f_L(\cdot; \mathbf{P}_{\le k}\mathbf{W})\|_{L^2}^2 \lesssim 1, \\
\mathcal{R}_3(\hat{\mathbf{W}}) &= \mathbb{E}_n\left[(f_L(\mathbf{x}; \mathbf{P}_{>k}\mathbf{W}))^2\right] \lesssim m^{-\frac{1}{2}} d^{\frac{k-1}{2}}, \\
\mathcal{R}_4(\hat{\mathbf{W}}) &= \|\mathbf{W}\|_{2,4}^8 \lesssim d^{2(k-1)}.
\end{aligned}$$

Therefore

$$L(\hat{\mathbf{W}}) \leq \mathbb{E}_\mu[\ell(y, f_Q(\mathbf{x}; \mathbf{W}) + f_L(\mathbf{x}; \mathbf{P}_{\leq k}\mathbf{W}))] + C \cdot m^{-\frac{1}{4}} d^{\frac{3(k-1)}{4}} + m^{-\frac{1}{4}} d^{\frac{k-1}{4}},$$

and similarly

$$\hat{L}(\hat{\mathbf{W}}) \geq \mathbb{E}_n[\ell(y, f_Q(\mathbf{x}; \mathbf{W}) + f_L(\mathbf{x}; \mathbf{P}_{\leq k}\mathbf{W}))] - C \cdot m^{-\frac{1}{4}} d^{\frac{3(k-1)}{4}} - m^{-\frac{1}{4}} d^{\frac{k-1}{4}}.$$

Since $\hat{L}(\hat{\mathbf{W}}) \leq C\varepsilon_{min}$, we thus have that

$$\mathbb{E}_n[\ell(y, f_Q(\mathbf{x}; \mathbf{W}) + f_L(\mathbf{x}; \mathbf{P}_{\leq k}\mathbf{W}))] \lesssim \varepsilon_{min} + m^{-\frac{1}{4}} d^{\frac{3(k-1)}{4}}.$$

Next, define the set

$$\mathcal{W} := \{\mathbf{W} \in \mathbb{R}^{m \times d} : \|\mathbf{W}\|_{2,4} \leq C d^{\frac{k-1}{4}}, \|f_L(\cdot; \mathbf{P}_{\leq k}\mathbf{W})\|_{L^2}^2 \leq C\}.$$

By construction, we have $\hat{\mathbf{W}} \in \mathcal{W}$. Furthermore, define the function class

$$\mathcal{L} := \{(\mathbf{x}, y) \to \ell(y, f_Q(\mathbf{x}; \mathbf{W}) + f_L(\mathbf{x}; \mathbf{P}_{\leq k}\mathbf{W})) : \mathbf{W} \in \mathcal{W}\}.$$

By the standard empirical Rademacher complexity bound, with probability $1 - d^{-9}$ over the draw of $\mathcal{D}$,

$$\mathbb{E}_\mu[\ell(y, f_Q(\mathbf{x}; \hat{\mathbf{W}}) + f_L(\mathbf{x}; \mathbf{P}_{\leq k}\hat{\mathbf{W}}))] - \mathbb{E}_n[\ell(y, f_Q(\mathbf{x}; \hat{\mathbf{W}}) + f_L(\mathbf{x}; \mathbf{P}_{\leq k}\hat{\mathbf{W}}))]] \leq 2\mathcal{R}_\mathcal{D}(\mathcal{L}),$$

By the Rademacher contraction Lemma [43], since ℓ is 1-Lipschitz we can bound (conditioning on Lemmas 35, 36)

$$\begin{aligned}
\mathcal{R}_\mathcal{D}(\mathcal{L}) &\leq \mathcal{R}_\mathcal{D}(\mathcal{F}_k^L(\mathcal{W}) + \mathcal{F}^Q(\mathcal{W})) + n^{-\frac{1}{2}} \\
&\leq \mathcal{R}_\mathcal{D}(\mathcal{F}_k^L(\mathcal{W})) + \mathcal{R}_\mathcal{D}(\mathcal{F}^Q(\mathcal{W})) + n^{-\frac{1}{2}} \\
&\lesssim \sqrt{\frac{d^k}{n}} \cdot \sup_{\mathbf{W} \in \mathcal{W}} \|f_L(\mathbf{x}; \mathbf{P}_{\leq k}\mathbf{W})\|_{L^2}^2 + \sqrt{\frac{d}{mn}} \cdot \sup_{\mathbf{W} \in \mathcal{W}} \|\mathbf{W}\|_F^2 + n^{-\frac{1}{2}} \\
&\lesssim \sqrt{\frac{d^k}{n}} + \sqrt{\frac{d}{n}} \cdot \sup_{\mathbf{W} \in \mathcal{W}} \|\mathbf{W}\|_{2,4}^2 + n^{-\frac{1}{2}} \\
&\lesssim \sqrt{\frac{d^k}{n}} + \sqrt{\frac{d}{n}} d^{\frac{k-1}{2}} + n^{-\frac{1}{2}} \\
&\lesssim \sqrt{\frac{d^k}{n}}.
\end{aligned}$$

Union bounding over the high probability events, with probability $1 - 3d^{-9} \geq 1 - d^{-8}$, we have

$$\begin{aligned}
L(\hat{\mathbf{W}}) &\leq \mathbb{E}_\mu[\ell(y, f_Q(\mathbf{x}; \hat{\mathbf{W}}) + f_L(\mathbf{x}; \mathbf{P}_{\leq k}\hat{\mathbf{W}}))] + C \cdot m^{-\frac{1}{4}} d^{\frac{3(k-1)}{4}} \\
&\leq \mathbb{E}_n[\ell(y, f_Q(\mathbf{x}; \hat{\mathbf{W}}) + f_L(\mathbf{x}; \mathbf{P}_{\leq k}\hat{\mathbf{W}}))]] + 2\mathcal{R}_\mathcal{D}(\mathcal{L}) + C \cdot m^{-\frac{1}{4}} d^{\frac{3(k-1)}{4}} \\
&\lesssim \varepsilon_{min} + m^{-\frac{1}{4}} d^{\frac{3(k-1)}{4}} + \sqrt{\frac{d^k}{n}} \\
&\lesssim \varepsilon_{min} + \sqrt{\frac{d^k}{n}}.
\end{aligned}$$

since $m \gtrsim d^{3(k-1)}\varepsilon_{min}^{-4}$. □

E.2 Proof of Lemma 35

Proof. Recall that we can write $f_L(\mathbf{x}; \mathbf{W}) = \varphi(\mathbf{x})^T \text{vec}(\mathbf{W})$, where $\varphi(\mathbf{x})$ is the NTK featurization map. Also, recall

$$\mathbf{\Sigma}_{\leq n_k} := \mathbb{E}_\mu\left[\varphi(\mathbf{x})^T \mathbf{P}_{\leq k} \varphi(\mathbf{x})\right] = \sum_{i=1}^{n_k} \lambda_i \mathbf{v}_i \mathbf{v}_i^T.$$

We can then bound

$$\begin{aligned}\mathcal{R}_{\mathcal{D}}(\mathcal{F}_k^L(\mathcal{W})) &= \mathbb{E}_\sigma\left[\sup_{\mathbf{W}\in\mathcal{W}}\frac{1}{n}\sum_{i=1}^n \sigma_i f_L(\mathbf{x};\mathbf{P}_{\le k}\mathrm{vec}(\mathbf{W}))\right]\\ &= \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{\mathbf{W}\in\mathcal{W}}\sum_{i=1}^n \sigma_i \varphi(\mathbf{x})^T\mathbf{P}_{\le k}\mathrm{vec}(\mathbf{W})\right]\\ &= \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{\mathbf{W}\in\mathcal{W}}\sum_{i=1}^n \sigma_i \varphi(\mathbf{x})^T\mathbf{P}_{\le k}(\mathbf{\Sigma}^\dagger_{\le n_k})^{1/2}\mathbf{\Sigma}^{1/2}_{\le n_k}\mathrm{vec}(\mathbf{W})\right]\\ &\le \frac{1}{n}\sup_{\mathbf{W}\in\mathcal{W}}\|\mathrm{vec}(\mathbf{W})\|_{\mathbf{\Sigma}_{\le n_k}}\mathbb{E}_\sigma\left\|\sum_{i=1}^n\sigma_i\mathbf{P}_{\le k}\varphi(\mathbf{x})\right\|_{\mathbf{\Sigma}^\dagger_{\le n_k}},\end{aligned}$$

where the last step follows by Cauchy-Schwarz. By definition,

$$\|\mathrm{vec}(\mathbf{W})\|_{\mathbf{\Sigma}_{\le n_k}} = \mathrm{vec}(\mathbf{W})^T\mathbf{\Sigma}_{\le n_k}\mathrm{vec}(\mathbf{W}) = \|f_L(\mathbf{x};\mathbf{P}_{\le k}\mathbf{W})\|_{L^2}^2.$$

Also, we can bound

$$\begin{aligned}\mathbb{E}_\sigma\left\|\sum_{i=1}^n\sigma_i\mathbf{P}_{\le k}\varphi(\mathbf{x})\right\|_{\mathbf{\Sigma}^\dagger_{\le n_k}} &\le \left(\mathbb{E}_\sigma\left\|\sum_{i=1}^n\sigma_i\mathbf{P}_{\le k}\varphi(\mathbf{x})\right\|^2_{\mathbf{\Sigma}^\dagger_{\le n_k}}\right)^{1/2}\\ &= (n\mathbb{E}_n[\varphi(\mathbf{x})^T\mathbf{P}_{\le k}\mathbf{\Sigma}^\dagger_{\le n_k}\mathbf{P}_{\le k}\varphi(\mathbf{x})])^{1/2}\\ &= \left(n\mathrm{Tr}\left(\mathbb{E}_n\left[(\mathbf{\Sigma}^\dagger_{\le n_k})^{1/2}\mathbf{P}_{\le k}\varphi(\mathbf{x})\varphi(\mathbf{x})^T\mathbf{P}_{\le k}(\mathbf{\Sigma}^\dagger_{\le n_k})^{1/2}\right]\right)\right)^{1/2}\end{aligned}$$

By Lemma 18, with probability $1-d^{-9}$ we have

$$\left\|\mathbb{E}_n\left[(\mathbf{\Sigma}^\dagger_{\le n_k})^{1/2}\mathbf{P}_{\le k}\varphi(\mathbf{x})\varphi(\mathbf{x})^T\mathbf{P}_{\le k}(\mathbf{\Sigma}^\dagger_{\le n_k})^{1/2}\right] - \mathbf{P}_{\le k}\right\|_{op} \le \frac{1}{2}.$$

Furthermore, $\mathbb{E}_n\left[(\mathbf{\Sigma}^\dagger_{\le n_k})^{1/2}\mathbf{P}_{\le k}\varphi(\mathbf{x})\varphi(\mathbf{x})^T\mathbf{P}_{\le k}(\mathbf{\Sigma}^\dagger_{\le n_k})^{1/2}\right]$ and $\mathbf{P}_{\le k}$ have the same span, which is dimension $n_k = \Theta(d^k)$. Therefore

$$\mathrm{Tr}\left(\mathbb{E}_n\left[(\mathbf{\Sigma}^\dagger_{\le n_k})^{1/2}\mathbf{P}_{\le k}\varphi(\mathbf{x})\varphi(\mathbf{x})^T\mathbf{P}_{\le k}(\mathbf{\Sigma}^\dagger_{\le n_k})^{1/2}\right]\right) \le \Theta(d^k).$$

Altogether, we get the bound

$$\mathbb{E}_\sigma\left\|\sum_{i=1}^n\sigma_i\mathbf{P}_{\le k}\varphi(\mathbf{x})\right\|_{\mathbf{\Sigma}^\dagger_{\le n_k}} \lesssim \sqrt{nd^k},$$

so

$$\mathcal{R}_n(\mathcal{F}_k^L(\mathcal{W})) \lesssim \sqrt{\frac{d^k}{n}}\cdot\sup_{\mathbf{W}\in\mathcal{W}}\|f_L(\mathbf{x};\mathbf{P}_{\le k}\mathbf{W})\|_{L^2}^2.$$

□

F Proof of Theorem 1

Proof. Choose $m \gtrsim n^4 d^{\frac{26(k+1)}{3}}\varepsilon^{-22/3}$. Set the regularization parameters as $\lambda_1 = m^{1/2}d^{-\frac{k-1}{2}}\varepsilon_{min}, \lambda_2 = \varepsilon_{min}, \lambda_3 = m^{1/2}d^{-\frac{k-1}{2}}\varepsilon_{min}\eta, \lambda_4 = d^{-2(k-1)}\varepsilon_{min}$. Also, set $r = n_k$, $\eta = c\lambda_4^{-1}m^{-3}$ for small constant c, and $\sigma = \tilde{O}(m^{-4})$.

For $\nu = m^{-1/2}, \gamma = m^{-3/4}$, $\tilde{\varepsilon} = \min(\nu, \gamma^2 m^{-5/2}) = m^{-4}$. Therefore by Theorem 5, with probability $1-d^{-8}$ we reach a (ν,γ)-SOSP within $\mathscr{T} = \tilde{O}(m^3\varepsilon^{-2}) = \tilde{O}(m^{11})$ timesteps. Call this point $\hat{\mathbf{W}}$.

By Corollary 1, $L_\lambda(\hat{\mathbf{W}}) \le C\varepsilon_{min}$. Finally, by Theorem 4, with probability $1-d^{-8}$, $L(\hat{\mathbf{W}}) \le C\varepsilon_{min}$. Setting $\varepsilon_{min} = \varepsilon/C$, we get that $L(\hat{\mathbf{W}}) \le \epsilon$ with probability $1-d^{-7}$, as desired. □

G Additional Experiments

G.1 Additional Simulations

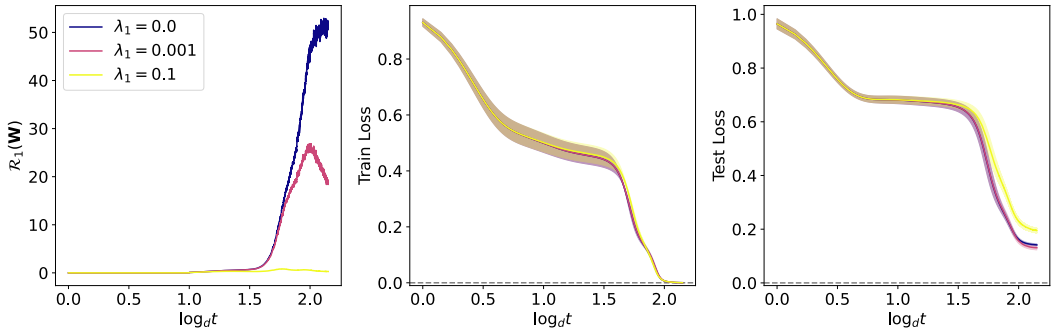

Figure 3: We train $f_L + f_Q$ with varying λ_1, while keeping λ_3 fixed

In Figure 3, we conduct the same experiment as in Figure 1, while additionally using the $\mathcal{R}_1$ regularizer. We fix $\lambda_3 = 0.01$ while varying λ_1. Since we cannot compute $\mathcal{R}_1$ exactly, we use an unbiased estimate at every timestep by sampling a new set of $\mathbf{x}$'s and computing $\mathbb{E}[f_L(\mathbf{x}; \mathbf{P}_{>k}\mathbf{W})]$ on this set. In the leftmost pane, we plot a moving average of our estimate of $\mathcal{R}_1$.

First, we observe that even when $\lambda_1 = 0$, the regularizer $\mathcal{R}_1$ is an order of magnitude smaller than $\mathcal{R}_3$ was when we set $\lambda_3 = 0$ (50 versus 500). Furthermore, in the rightmost pane, we see that the test loss of the model is small regardless of which value of λ_1 was chosen. This provides additional evidence that $\mathcal{R}_1$ is kept small throughout the training process.

G.2 CIFAR10 Experiments

To demonstrate the significance of our approach on "realistic" datasets/models, we consider experiments with CNNs on CIFAR10.

[8] showed that, in practice, training the second-order Taylor expansion of the network tracks the true gradient descent dynamics far better than the network's linearization does. This is further demonstrated in Figure 4. Here, we train a 4-layer CNN with width 512, ReLU activation, average pooling between each layer, and the standard PyTorch initialization, on the cats vs. horses CIFAR10 classification task. We train via SGD with batch size 128. For both the train loss and test loss, the dynamics from training $f_L + f_Q$ tracks the true network dynamics better than just the linearization f_L. Furthermore, $f_L + f_Q$ acheives a lower test loss than f_L (the true network beats both Taylorizations).

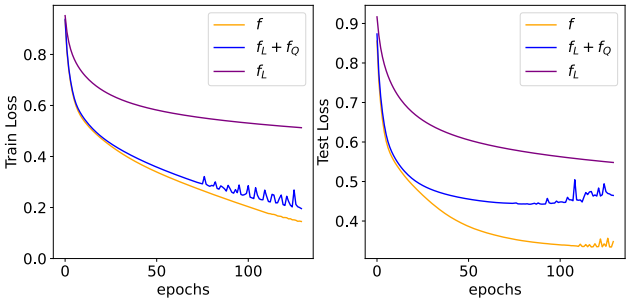

Figure 4: $f_L + f_Q$ tracks the true network dynamics far better than just f_L.

In Table 1, we additionally measure the test loss of the linear and quadratic terms *after 100 epochs of training the full model*. We observe that both the linear and quadratic term have a nontrivial loss (< 1), and thus learned a nonzero component of the signal. This provides evidence that in real neural networks, both the linear and quadratic terms learn a nontrivial component of the signal.

Model	f	$f_L + f_Q$	f_L
Test Loss	0.340	0.699	0.887
Test Accuracy	90.8%	76.7%	60.2%

Table 1: The test loss and accuracy of f, f_L, and $f_L + f_Q$ evaluated on the iterate obtained after 100 epochs of training using the model f.

G.3 Standard MLP Experimental Details

In Figure 2, we demonstrated that "standard" neural networks can effectively learn low-degree dense and high-degree sparse polynomials. We trained a 2-layer neural network with standard PyTorch initialization and width 100 to learn the target function $f^*(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x} + h_3(\beta^T\mathbf{x})$, where $\mathbf{A}$ is a high-rank matrix chosen so that $\mathbf{x} \to \mathbf{x}^T\mathbf{A}\mathbf{x}$ has an L^2 norm of 1. Here, h_3 is the 3rd Hermite polynomial, and thus $h_3(\beta^T\mathbf{x})$ is a sparse cubic only depending on the random direction β (the Hermite polynomial is chosen for this task so that it is orthogonal to the quadratic term, making it the "hardest" low-rank cubic to learn).

For varying values of dimension d from 10 to 100 and number of samples n, we train our network via vanilla gradient descent with fixed learning rate 0.05. The initialization, small width, fixed learning rate, and lack of regularization are designed to mimic a standard deep learning setup. For each value of d, we compute the minimum n required such that the test loss is < 0.1 (note that the test loss of the zero predictor is 1.0). Figure 2 is a $\log - \log$ plot of d versus this optimal n.

In Figure 2, we observe that the number of samples needed to obtain 0.1 test loss roughly scales with d^2. We convincingly see that much fewer than d^3 samples (the red dashed line) are needed. The NTK, on the otherhand, requires $\Omega(d^3)$ samples to learn any cubic function. The minimax sample complexity to learn arbitrary quadratics is $\Theta(d^2)$, and therefore this experiment shows that standard neural networks learn "dense qudratic plus sparse cubic" functions with optimal sample complexity. This provides further evidence that the low-degree plus sparse task is worthy of theoretical study.

Experimental Details. All experiments were run on an NVIDIA RTX A6000 GPU. We use the JAX framework [10] along with the Neural Tangents API [40]. Code for all experiments can be found at `https://github.com/eshnich/escape_NTK`.