Understanding Why Generalized Reweighting Does Not Improve Over ERM (Appendix)

Anonymous Author(s)
Affiliation
Address
email

A Related work

Group Fairness. Group fairness in machine learning was first studied in [HPS16] and [ZVGRG17], where they required the model to perform equally well over all groups. Later, [HSNL18] studied another type of group fairness called Rawlsian max-min fairness [Raw01], which does not require equal performance but rather requires high performance on the worst-off group. The subpopulation shift problem we study in this paper is most closely related to Rawlsian max-min fairness. A large body of recent work have studied how to improve this worst-group performance [DN18, OSHL19, LHC$^+$21, ZDKR21]. Recent work however observe that these approaches, when used with modern overparameterized models, easily overfit [SKHL20, SRKL20]. Apart from group fairness, there are also other notions of fairness, such as individual fairness [DHP$^+$12, ZWS$^+$13] and counterfactual fairness [KLRS17], which we do not study in this work.

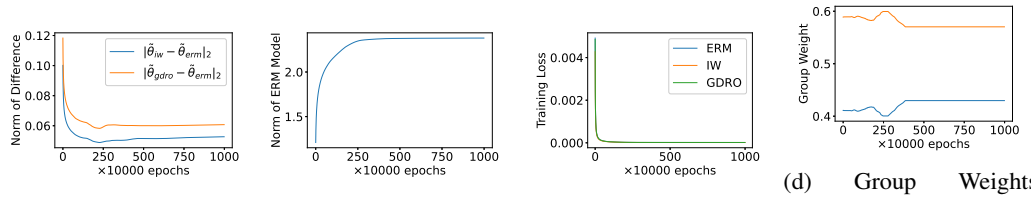
Implicit Bias Under the Overparameterized Setting. For overparameterized models, there could be many model parameters which all minimize the training loss. In such cases, it is of interest to study the implicit bias of specific optimization algorithms such as gradient descent i.e. to what minimizer the model parameters will converge to [DLL$^+$19, AZLS19]. Our results use the NTK formulation of wide neural networks [JGH18], and specifically we use linearized neural networks to approximate such wide neural networks following [LXS$^+$19]. There is some criticism of this line of work, e.g. [COB19] argued that infinitely wide neural networks fall in the "lazy training" regime and results might not be transferable to general neural networks. Nonetheless such wide neural networks are being widely studied in recent years, since they provide considerable insights into the behavior of more general neural networks, which are typically intractable to analyze otherwise.

B Extension to Multi-dimensional Regression / Multi-class Classification

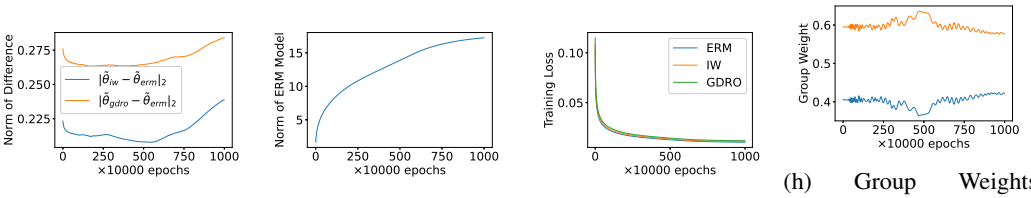
In our results, we assume that $f : \mathbb{R}^d \to \mathbb{R}$ for simplicity, but our results can be very easily extended to the case where $f : \mathbb{R}^d \to \mathbb{R}^k$. For most of our results, the proof consists of two major components: (i) The linearized neural network will converge to some point (interpolator, max-margin classifier, etc.); (ii) The wide fully-connected neural network can be approximated by its linearized counterpart. For both components, the extension is very simple and straightforward. For (i), the proof only relies on the smoothness of the objective function and the upper quadratic bound it entails, and the function is still smooth when its output becomes multi-dimensional; For (ii), we can prove that $\sup_t \|f(\boldsymbol{x}) - f_{\text{lin}}(\boldsymbol{x})\|_2 = O(\tilde{d}^{-1/4})$ in exactly the same way. Thus, all of our results hold for multi-dimensional regression and multi-class classification.

Particularly, for the multi-class cross-entropy loss, using Theorem 8 we can show that under any GRW satisfying Assumption 1, the direction of the weight of a linear classifier will converge to the

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.



(a) Weight Difference (b) Norm of ERM Model (c) Training Loss (d) Group Weights (GDRO)



(e) Weight Difference (f) Norm of ERM Model (g) Training Loss (h) Group Weights (GDRO)

Figure 1: Experimental results of ERM, importance weighting (IW) and Group DRO (GDRO) with the logistic loss and the polynomially-tailed loss. First row: Logistic loss; Second row: Polynomially-tailed loss. All norms are L_2 norms. $\tilde{\theta}$ is a unit vector which is the direction of θ.

following max-margin classifier:

$$\hat{\theta}_{\text{MM}} = \arg\min_{\theta} \left\{ \min_{i=1,\cdots,n} \left[f(\boldsymbol{x}_i)_{y_i} - \max_{y' \neq y_i} f(\boldsymbol{x}_i)_{y'} \right] : \|\theta\|_2 = 1 \right\} \tag{20}$$

which is still independent of q_i.

C More Experiments

We run ERM, importance weighting and Group DRO on the training set with 6 MNIST images which we used in Section 4.1 with the logistic loss and the polynomially-tailed loss (Eqn. (19), with $\alpha = 1$, $\beta = 0$ and ℓ_{left} being the logistic loss shifted to make the overall loss function continuous) on this dataset for 10 million epochs (note that we run for much more epochs because the convergence is very slow). The results are shown in Figure 1. From the plots we can see that:

- For either loss function, the training loss of each method converges to 0.
- In contrast to the theory that the norm of the ERM model will go to infinity and all models will converge to the max-margin classifier, the weight of the ERM model gets stuck at some point, and the norms of the gaps between the normalized model weights also get stuck. The reason is that the training loss has got so small that it becomes zero in the floating number representation, so the gradient also becomes zero and the training halts due to limited computational precision.
- However, we can still observe a fundamental difference between the logistic loss and the polynomially-tailed loss. For the logistic loss, the norm of the gap between importance weighting (or Group DRO) and ERM will converge to around 0.06 when the training stops, while for the polynomially-tailed loss, the norm will be larger than 0.22 and will keep growing, which shows that for the polynomially-tailed loss the normalized model weights do not converge to the same point.
- For either loss, the group weights of Group DRO still empirically satisfy Assumption 1.

D Proofs

In this paper, for any matrix $\boldsymbol{A}$, we will use $\|\boldsymbol{A}\|_2$ to denote its spectral norm and $\|\boldsymbol{A}\|_F$ to denote its Frobenius norm.

D.1 Background on Smoothness

A first-order differentiable function f over $\mathcal{D}$ is called *L-smooth* for $L > 0$ if

$$f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{L}{2} \|\boldsymbol{y} - \boldsymbol{x}\|_2^2 \qquad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{D} \tag{21}$$

which is also called the *upper quadratic bound*. If f is second-order differentiable and $\mathcal{D}$ is a convex set, then f is L-smooth is equivalent to

$$\boldsymbol{v}^\top \nabla^2 f(\boldsymbol{x}) \boldsymbol{v} \leq L \qquad \forall \|\boldsymbol{v}\|_2 = 1, \forall \boldsymbol{x} \in \mathcal{D} \tag{22}$$

A classical result in convex optimization is the following:

Theorem 10. *If $f(\boldsymbol{x})$ is convex and L-smooth with a unique finite minimizer $\boldsymbol{x}^*$, and is minimized by gradient descent $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta \nabla f(\boldsymbol{x}_t)$ starting from $\boldsymbol{x}_0$ where the learning rate $\eta \leq \frac{1}{L}$, then we have*

$$f(\boldsymbol{x}_T) \leq f(\boldsymbol{x}^*) + \frac{1}{\eta T} \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2 \tag{23}$$

which also implies that $\boldsymbol{x}_T$ converges to $\boldsymbol{x}^*$ as $T \to \infty$.

D.2 Proofs for Subsection 4.1

D.2.1 Proof of Theorem 1

Using the key intuition, the weight update rule (7) implies that $\theta^{(t+1)} - \theta^{(t)} \in \text{span}\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$ for all t, which further implies that $\theta^{(t)} - \theta^{(0)} \in \text{span}\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$ for all t. By Cramer's rule, in this n-dimensional subspace there exists one and only one θ^* such that $\theta^* - \theta^{(0)} \in \text{span}\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$ and $\langle \theta^*, \boldsymbol{x}_i \rangle$ for all i. Then we have

$$\left\| \boldsymbol{X}^\top (\theta^{(t)} - \theta^*) \right\|_2 = \left\| (\boldsymbol{X}^\top \theta^{(t)} - \boldsymbol{Y}) - (\boldsymbol{X}^\top \theta^* - \boldsymbol{Y}) \right\|_2 \leq \left\| \boldsymbol{X}^\top \theta^{(t)} - \boldsymbol{Y} \right\|_2 + \left\| \boldsymbol{X}^\top \theta^* - \boldsymbol{Y} \right\|_2 \to 0 \tag{24}$$

because $\left\| \boldsymbol{X}^\top \theta - \boldsymbol{Y} \right\|_2^2 = 2n\hat{\mathcal{R}}(f(\boldsymbol{x}; \theta))$. On the other hand, let $s^{\min}$ be the smallest singular value of $\boldsymbol{X}$. Since $\boldsymbol{X}$ is full-rank, $s^{\min} > 0$, and $\left\| \boldsymbol{X}^\top (\theta^{(t)} - \theta^*) \right\|_2 \geq s^{\min} \left\| \theta^{(t)} - \theta^* \right\|_2$. This shows that $\left\| \theta^{(t)} - \theta^* \right\|_2 \to 0$. Thus, $\theta^{(t)}$ converges to this unique θ^*. □

D.2.2 Proof of Theorem 2

To help our readers understand the proof more easily, we will first prove the result for static GRW where $q_i^{(t)} = q_i$ for all t, and then we will prove the result for dynamic GRW that satisfy $q_i^{(t)} \to q_i$ as $t \to \infty$.

Static GRW. We first prove the result for all static GRW such that $\min_i q_i = q^* > 0$.

We will use smoothness introduce in Appendix D.1. Denote $A = \sum_{i=1}^n \|\boldsymbol{x}_i\|_2^2$. The empirical risk of the linear model $f(\boldsymbol{x}) = \langle \theta, \boldsymbol{x} \rangle$ is

$$F(\theta) = \sum_{i=1}^n q_i (\boldsymbol{x}_i^\top \theta - y_i)^2 \tag{25}$$

whose Hessian is

$$\nabla_\theta^2 F(\theta) = 2 \sum_{i=1}^n q_i \boldsymbol{x}_i \boldsymbol{x}_i^\top \tag{26}$$

So for any unit vector $\boldsymbol{v} \in \mathbb{R}^d$, we have (since $q_i \in [0, 1]$)

$$\boldsymbol{v}^\top \nabla_\theta^2 F(\theta) \boldsymbol{v} = 2 \sum_{i=1}^n q_i (\boldsymbol{x}_i^\top \boldsymbol{v})^2 \leq 2 \sum_{i=1}^n q_i \|\boldsymbol{x}_i\|_2^2 \leq 2A \tag{27}$$

which implies that $F(\theta)$ is $2A$-smooth. Thus, we have the following upper quadratic bound: for any $\theta_1, \theta_2 \in \mathbb{R}^d$,

$$F(\theta_2) \leq F(\theta_1) + \langle \nabla_\theta F(\theta_1), \theta_2 - \theta_1 \rangle + A \left\| \theta_2 - \theta_1 \right\|_2^2 \tag{28}$$

Denote $g(\theta^{(t)}) = (\boldsymbol{X}^\top \theta^{(t)} - \boldsymbol{Y}) \in \mathbb{R}^n$. We can see that $\left\| \sqrt{\boldsymbol{Q}} g(\theta^{(t)}) \right\|_2^2 = F(\theta^{(t)})$, where where $\sqrt{\boldsymbol{Q}} = \text{diag}(\sqrt{q_1}, \cdots, \sqrt{q_n})$. Thus, $\nabla F(\theta^{(t)}) = 2\boldsymbol{X}\boldsymbol{Q}g(\theta^{(t)})$. The update rule of a static GRW with gradient descent and the squared loss is:

$$\theta^{(t+1)} = \theta^{(t)} - \eta \sum_{i=1}^n q_i \boldsymbol{x}_i (f^{(t)}(\boldsymbol{x}_i) - y_i) = \theta^{(t)} - \eta \boldsymbol{X}\boldsymbol{Q}g(\theta^{(t)}) \tag{29}$$

Substituting θ_1 and θ_2 in (28) with $\theta^{(t)}$ and $\theta^{(t+1)}$ yields

$$F(\theta^{(t+1)}) \leq F(\theta^{(t)}) - 2\eta g(\theta^{(t)})^\top \boldsymbol{Q}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{Q}g(\theta^{(t)}) + A \left\| \eta \boldsymbol{X}\boldsymbol{Q}g(\theta^{(t)}) \right\|_2^2 \tag{30}$$

Since $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$ are linearly independent, $\boldsymbol{X}^\top \boldsymbol{X}$ is a positive definite matrix. Denote the smallest eigenvalue of $\boldsymbol{X}^\top \boldsymbol{X}$ by $\lambda^{\min} > 0$. And $\left\| \boldsymbol{Q}g(\theta^{(t)}) \right\|_2 \geq \sqrt{q^*} \left\| g(\theta^{(t)}) \right\|_2 = \sqrt{q^* F(\theta^{(t)})}$, so we have $g(\theta^{(t)})^\top \boldsymbol{Q}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{Q}g(\theta^{(t)}) \geq q^* \lambda^{\min} F(\theta^{(t)})$. Thus,

$$\begin{aligned}
F(\theta^{(t+1)}) &\leq F(\theta^{(t)}) - 2\eta q^* \lambda^{\min} F(\theta^{(t)}) + A\eta^2 \left\| \boldsymbol{X}\sqrt{\boldsymbol{Q}} \right\|_2^2 \left\| \sqrt{\boldsymbol{Q}} g(\theta^{(t)}) \right\|_2^2 \\
&\leq F(\theta^{(t)}) - 2\eta q^* \lambda^{\min} F(\theta^{(t)}) + A\eta^2 \left\| \boldsymbol{X}\sqrt{\boldsymbol{Q}} \right\|_F^2 F(\theta^{(t)}) \\
&\leq F(\theta^{(t)}) - 2\eta q^* \lambda^{\min} F(\theta^{(t)}) + A\eta^2 \left\| \boldsymbol{X} \right\|_F^2 F(\theta^{(t)}) \\
&= (1 - 2\eta q^* \lambda^{\min} + A^2\eta^2) F(\theta^{(t)})
\end{aligned} \tag{31}$$

Let $\eta_0 = \frac{q^* \lambda^{\min}}{A^2}$. For any $\eta \leq \eta_0$, we have $F(\theta^{(t+1)}) \leq (1 - \eta q^* \lambda^{\min}) F(\theta^{(t)})$ for all t, which implies that $\lim_{t \to \infty} F(\theta^{(t)}) = 0$. This implies that the empirical training risk must converge to 0.

Dynamic GRW. Now we prove the result for all dynamic GRW satisfying Assumption 1. By Assumption 1, for any $\epsilon > 0$, there exists t_ϵ such that for all $t \geq t_\epsilon$ and all i,

$$q_i^{(t)} \in (q_i - \epsilon, q_i + \epsilon) \tag{32}$$

This is because for all i, there exists t_i such that for all $t \geq t_i$, $q_i^{(t)} \in (q_i - \epsilon, q_i + \epsilon)$. Then, we can define $t_\epsilon = \max\{t_1, \cdots, t_n\}$. Denote the largest and smallest eigenvalues of $\boldsymbol{X}^\top \boldsymbol{X}$ by $\lambda^{\max}$ and $\lambda^{\min}$, and because $\boldsymbol{X}$ is full-rank, we have $\lambda^{\min} > 0$. Define $\epsilon = \min\{\frac{q^*}{3}, \frac{(q^* \lambda^{\min})^2}{12 \lambda^{\max 2}}\}$, and then t_ϵ is also fixed.

We still denote $\boldsymbol{Q} = \text{diag}(q_1, \cdots, q_n)$. When $t \geq t_\epsilon$, the update rule of a dynamic GRW with gradient descent and the squared loss is:

$$\theta^{(t+1)} = \theta^{(t)} - \eta \boldsymbol{X}\boldsymbol{Q}_\epsilon^{(t)} (\boldsymbol{X}^\top \theta^{(t)} - \boldsymbol{Y}) \tag{33}$$

where $\boldsymbol{Q}_\epsilon^{(t)} = \boldsymbol{Q}^{(t)}$, and we use the subscript ϵ to indicate that $\left\| \boldsymbol{Q}_\epsilon^{(t)} - \boldsymbol{Q} \right\|_2 < \epsilon$. Then, note that we can rewrite $\boldsymbol{Q}_\epsilon^{(t)}$ as $\boldsymbol{Q}_\epsilon^{(t)} = \sqrt{\boldsymbol{Q}_{3\epsilon}^{(t)}} \cdot \sqrt{\boldsymbol{Q}}$ as long as $\epsilon \leq q^*/3$. This is because $q_i + \epsilon \leq \sqrt{(q_i + 3\epsilon) q_i}$ and $q_i - \epsilon \geq \sqrt{(q_i - 3\epsilon) q_i}$ for all $\epsilon \leq q_i/3$, and $q_i \geq q^*$. Thus, we have

$$\theta^{(t+1)} = \theta^{(t)} - \eta \boldsymbol{X} \sqrt{\boldsymbol{Q}_{3\epsilon}^{(t)}} \sqrt{\boldsymbol{Q}} g(\theta^{(t)}) \quad \text{where } \boldsymbol{Q}_\epsilon^{(t)} = \sqrt{\boldsymbol{Q}_{3\epsilon}^{(t)}} \cdot \sqrt{\boldsymbol{Q}} \tag{34}$$

Again, substituting θ_1 and θ_2 in (28) with $\theta^{(t)}$ and $\theta^{(t+1)}$ yields

$$F(\theta^{(t+1)}) \leq F(\theta^{(t)}) - 2\eta g(\theta^{(t)})^\top \boldsymbol{Q}^\top \boldsymbol{X}^\top \boldsymbol{X} \sqrt{\boldsymbol{Q}_{3\epsilon}^{(t)}} \sqrt{\boldsymbol{Q}} g(\theta^{(t)}) + A \left\| \eta \boldsymbol{X} \sqrt{\boldsymbol{Q}_{3\epsilon}^{(t)}} \sqrt{\boldsymbol{Q}} g(\theta^{(t)}) \right\|_2^2 \tag{35}$$

Then, note that

$$\begin{aligned}
&\left|g(\theta^{(t)})^\top \boldsymbol{Q}^\top \boldsymbol{X}^\top \boldsymbol{X}\left(\sqrt{\boldsymbol{Q}_{3\epsilon}^{(t)}}-\sqrt{\boldsymbol{Q}}\right)\sqrt{\boldsymbol{Q}}g(\theta^{(t)})\right| \\
\leq&\left\|\sqrt{\boldsymbol{Q}}^\top \boldsymbol{X}^\top \boldsymbol{X}\left(\sqrt{\boldsymbol{Q}_{3\epsilon}^{(t)}}-\sqrt{\boldsymbol{Q}}\right)\right\|_2\left\|\sqrt{\boldsymbol{Q}}g(\theta^{(t)})\right\|_2^2 \\
\leq&\left\|\sqrt{\boldsymbol{Q}}\right\|_2\left\|\boldsymbol{X}^\top \boldsymbol{X}\right\|_2\left\|\sqrt{\boldsymbol{Q}_{3\epsilon}^{(t)}}-\sqrt{\boldsymbol{Q}}\right\|_2\left\|\sqrt{\boldsymbol{Q}}g(\theta^{(t)})\right\|_2^2 \\
\leq&\lambda^{\max}\sqrt{3\epsilon}F(\theta^{(t)})
\end{aligned} \tag{36}$$

where the last step comes from the following fact: for all $\epsilon < q_i/3$,

$$\sqrt{q_i+3\epsilon}-\sqrt{q_i}\leq\sqrt{3\epsilon}\quad\text{and}\quad\sqrt{q_i}-\sqrt{q_i-3\epsilon}\leq\sqrt{3\epsilon} \tag{37}$$

And as proved before, we also have

$$g(\theta^{(t)})^\top \boldsymbol{Q}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{Q}g(\theta^{(t)})\geq q^*\lambda^{\min}F(\theta^{(t)}) \tag{38}$$

Since $\epsilon\leq\frac{(q^*\lambda^{\min})^2}{12\lambda^{\max 2}}$, we have

$$g(\theta^{(t)})^\top \boldsymbol{Q}^\top \boldsymbol{X}^\top \boldsymbol{X}\sqrt{\boldsymbol{Q}_{3\epsilon}^{(t)}}\sqrt{\boldsymbol{Q}}g(\theta^{(t)})\geq\left(q^*\lambda^{\min}-\lambda^{\max}\sqrt{3\epsilon}\right)F(\theta^{(t)})\geq\frac{1}{2}q^*\lambda^{\min}F(\theta^{(t)}) \tag{39}$$

Thus,

$$\begin{aligned}
F(\theta^{(t+1)})&\leq F(\theta^{(t)})-\eta q^*\lambda^{\min}F(\theta^{(t)})+A\eta^2\left\|\boldsymbol{X}\sqrt{\boldsymbol{Q}_{3\epsilon}^{(t)}}\right\|_2^2\left\|\sqrt{\boldsymbol{Q}}g(\theta^{(t)})\right\|_2^2 \\
&\leq(1-\eta q^*\lambda^{\min}+A^2\eta^2(1+3\epsilon))F(\theta^{(t)}) \\
&\leq(1-\eta q^*\lambda^{\min}+2A^2\eta^2)F(\theta^{(t)})
\end{aligned} \tag{40}$$

for all $\epsilon\leq 1/3$. Let $\eta_0=\frac{q^*\lambda^{\min}}{4A^2}$. For any $\eta\leq\eta_0$, we have $F(\theta^{(t+1)})\leq(1-\eta q^*\lambda^{\min}/2)F(\theta^{(t)})$ for all $t\geq t_\epsilon$, which implies that $\lim_{t\to\infty}F(\theta^{(t)})=0$. Thus, the empirical training risk converges to 0. □

D.3 Proofs for Subsection 4.2

D.3.1 Proof of Lemma 3

Note that the first l layers (except the output layer) of the original NTK formulation and our new formulation are the same, so we still have the following proposition:

Proposition 11 (Proposition 1 in [JGH18]). *If σ is Lipschitz and $d_l\to\infty$ for $l=1,\cdots,L$ sequentially, then for all $l=1,\cdots,L$, the distribution of a single element of $\boldsymbol{h}^l$ converges in probability to a zero-mean Gaussian process of covariance Σ^l that is defined recursively by:*

$$\begin{aligned}
\Sigma^1(\boldsymbol{x},\boldsymbol{x}')&=\frac{1}{d_0}\boldsymbol{x}^\top\boldsymbol{x}'+\beta^2 \\
\Sigma^l(\boldsymbol{x},\boldsymbol{x}')&=\mathbb{E}_f[\sigma(f(\boldsymbol{x}))\sigma(f(\boldsymbol{x}'))]+\beta^2
\end{aligned} \tag{41}$$

where f is sampled from a zero-mean Gaussian process of covariance $\Sigma^{(l-1)}$.

Now we show that for an infinitely wide neural network with $L\geq 1$ hidden layers, $\Theta^{(0)}$ converges in probability to the following non-degenerated deterministic limiting kernel

$$\Theta=\mathbb{E}_{f\sim\Sigma^L}[\sigma(f(\boldsymbol{x}))\sigma(f(\boldsymbol{x}'))]+\beta^2 \tag{42}$$

Consider the output layer $\boldsymbol{h}^{L+1}=\frac{W^L}{\sqrt{\tilde{d}}}\sigma(\boldsymbol{h}^L)+\beta\boldsymbol{b}^L$. We can see that for any parameter θ_i before the output layer,

$$\nabla_{\theta_i}\boldsymbol{h}^{L+1}=\text{diag}(\dot{\sigma}(\boldsymbol{h}^L))\frac{W^{L\top}}{\sqrt{d_L}}\nabla_{\theta_i}\boldsymbol{h}^L=0 \tag{43}$$

And for W^L and $\boldsymbol{b}^L$, we have

$$\nabla_{W^L}\boldsymbol{h}^{L+1} = \frac{1}{\sqrt{d_L}}\sigma(\boldsymbol{h}^L) \qquad \text{and} \qquad \nabla_{\boldsymbol{b}^L}\boldsymbol{h}^{L+1} = \beta \tag{44}$$

Then we can achieve (42) by the law of large numbers. □

D.3.2 Proof of Lemma 5

We will use the following short-hand in the proof:

$$\begin{cases} g(\theta^{(t)}) = f^{(t)}(\boldsymbol{X}) - \boldsymbol{Y} \\ J(\theta^{(t)}) = \nabla_\theta f(\boldsymbol{X};\theta^{(t)}) \in \mathbb{R}^{p\times n} \\ \Theta^{(t)} = J(\theta^{(t)})^\top J(\theta^{(t)}) \end{cases} \tag{45}$$

For any $\epsilon > 0$, there exists t_ϵ such that for all $t \geq t_\epsilon$ and all i, $q_i^{(t)} \in (q_i - \epsilon, q_i + \epsilon)$. Like what we have done in (34), we can rewrite $\boldsymbol{Q}^{(t)} = \boldsymbol{Q}_\epsilon^{(t)} = \sqrt{\boldsymbol{Q}_{3\epsilon}^{(t)}} \cdot \sqrt{\boldsymbol{Q}}$, where $\boldsymbol{Q} = \text{diag}(q_1, \cdots, q_n)$.

The update rule of a GRW with gradient descent and the squared loss for the wide neural network is:

$$\theta^{(t+1)} = \theta^{(t)} - \eta J(\theta^{(t)})\boldsymbol{Q}^{(t)} g(\theta^{(t)}) \tag{46}$$

and for $t \geq t_\epsilon$, it can be rewritten as

$$\theta^{(t+1)} = \theta^{(t)} - \eta J(\theta^{(t)})\sqrt{\boldsymbol{Q}_{3\epsilon}^{(t)}}\left[\sqrt{\boldsymbol{Q}} g(\theta^{(t)})\right] \tag{47}$$

First, we will prove the following theorem:

Theorem 12. *There exist constants $M > 0$ and $\epsilon_0 > 0$ such that for all $\epsilon \in (0, \epsilon_0]$, $\eta \leq \eta^*$ and any $\delta > 0$, there exist $R_0 > 0$, $\tilde{D} > 0$ and $B > 1$ such that for any $\tilde{d} \geq \tilde{D}$, the following (i) and (ii) hold with probability at least $(1-\delta)$ over random initialization when applying gradient descent with learning rate η:*

(i) For all $t \leq t_\epsilon$, there is

$$\left\|g(\theta^{(t)})\right\|_2 \leq B^t R_0 \tag{48}$$

$$\sum_{j=1}^{t}\left\|\theta^{(j)} - \theta^{(j-1)}\right\|_2 \leq \eta M R_0 \sum_{j=1}^{t} B^{j-1} < \frac{MB^{t_\epsilon}R_0}{B-1} \tag{49}$$

(ii) For all $t \geq t_\epsilon$, we have

$$\left\|\sqrt{\boldsymbol{Q}} g(\theta^{(t)})\right\|_2 \leq \left(1 - \frac{\eta q^* \lambda^{\min}}{3}\right)^{t-t_\epsilon} B^{t_\epsilon} R_0 \tag{50}$$

$$\begin{aligned} \sum_{j=t_\epsilon+1}^{t}\left\|\theta^{(j)} - \theta^{(j-1)}\right\|_2 &\leq \eta\sqrt{1+3\epsilon} M B^{t_\epsilon} R_0 \sum_{j=t_\epsilon+1}^{t}\left(1 - \frac{\eta q^*\lambda^{\min}}{3}\right)^{j-t_\epsilon} \\ &< \frac{3\sqrt{1+3\epsilon} M B^{t_\epsilon} R_0}{q^*\lambda^{\min}} \end{aligned} \tag{51}$$

Proof. The proof is based on the following lemma:

Lemma 13 (Local Lipschitzness of the Jacobian)**.** *Under Assumption 2, there is a constant $M > 0$ such that for any $C_0 > 0$ and any $\delta > 0$, there exists a $\tilde{D}$ such that: If $\tilde{d} \geq \tilde{D}$, then with probability at least $(1-\delta)$ over random initialization, for any $\boldsymbol{x}$ such that $\|\boldsymbol{x}\|_2 \leq 1$,*

$$\begin{cases} \left\|\nabla_\theta f(\boldsymbol{x};\theta) - \nabla_\theta f(\boldsymbol{x};\tilde{\theta})\right\|_2 \leq \frac{M}{\sqrt[4]{\tilde{d}}}\left\|\theta - \tilde{\theta}\right\|_2 \\ \|\nabla_\theta f(\boldsymbol{x};\theta)\|_2 \leq M \\ \left\|J(\theta) - J(\tilde{\theta})\right\|_F \leq \frac{M}{\sqrt[4]{\tilde{d}}}\left\|\theta - \tilde{\theta}\right\|_2 \\ \|J(\theta)\|_F \leq M \end{cases}, \qquad \forall \theta, \tilde{\theta} \in B(\theta^{(0)}, C_0) \tag{52}$$

where $B(\theta^{(0)}, R) = \{\theta : \left\|\theta - \theta^{(0)}\right\|_2 < R\}$.

The proof of this lemma can be found in Appendix D.3.3. Note that for any $\boldsymbol{x}$, $f^{(0)}(\boldsymbol{x}) = \beta \boldsymbol{b}^L$ where $\boldsymbol{b}^L$ is sampled from the standard Gaussian distribution. Thus, for any $\delta > 0$, there exists a constant R_0 such that with probability at least $(1 - \delta/3)$ over random initialization,

$$\left\|g(\theta^{(0)})\right\|_2 < R_0 \tag{53}$$

And by Proposition 3, there exists $D_2 \geq 0$ such that for any $\tilde{d} \geq D_2$, with probability at least $(1 - \delta/3)$,

$$\left\|\Theta - \Theta^{(0)}\right\|_F \leq \frac{q^* \lambda^{\min}}{3} \tag{54}$$

Let M be the constant in Lemma 13. Let $\epsilon_0 = \frac{(q^*\lambda^{\min})^2}{108M^4}$. Let $B = 1 + \eta^* M^2$, and $C_0 = \frac{MB^{t_\epsilon}R_0}{B-1} + \frac{3\sqrt{1+3\epsilon}MB^{t_\epsilon}R_0}{q^*\lambda^{\min}}$. By Lemma 13, there exists $D_1 > 0$ such that with probability at least $(1 - \delta/3)$, for any $\tilde{d} \geq D_1$, (52) is true for all $\theta, \tilde{\theta} \in B(\theta^{(0)}, C_0)$.

By union bound, with probability at least $(1 - \delta)$, (52), (53) and (54) are all true. Now we assume that all of them are true, and prove (48) and (49) by induction. (48) is true for $t = 0$ due to (53), and (49) is always true for $t = 0$. Suppose (48) and (49) are true for t, then for $t + 1$ we have

$$\begin{aligned}\left\|\theta^{(t+1)} - \theta^{(t)}\right\|_2 &\leq \eta \left\|J(\theta^{(t)})\boldsymbol{Q}^{(t)}\right\|_2 \left\|g(\theta^{(t)})\right\|_2 \leq \eta \left\|J(\theta^{(t)})\boldsymbol{Q}^{(t)}\right\|_F \left\|g(\theta^{(t)})\right\|_2 \\ &\leq \eta \left\|J(\theta^{(t)})\right\|_F \left\|g(\theta^{(t)})\right\|_2 \leq M\eta B^t R_0\end{aligned} \tag{55}$$

So (49) is also true for $t + 1$. And we also have

$$\begin{aligned}\left\|g(\theta^{(t+1)})\right\|_2 &= \left\|g(\theta^{(t+1)}) - g(\theta^{(t)}) + g(\theta^{(t)})\right\|_2 \\ &= \left\|J(\tilde{\theta}^{(t)})^\top(\theta^{(t+1)} - \theta^{(t)}) + g(\theta^{(t)})\right\|_2 \\ &= \left\|-\eta J(\tilde{\theta}^{(t)})^\top J(\theta^{(t)})\boldsymbol{Q}^{(t)} g(\theta^{(t)}) + g(\theta^{(t)})\right\|_2 \\ &\leq \left\|\boldsymbol{I} - \eta J(\tilde{\theta}^{(t)})^\top J(\theta^{(t)})\boldsymbol{Q}^{(t)}\right\|_2 \left\|g(\theta^{(t)})\right\|_2 \\ &\leq \left(1 + \left\|\eta J(\tilde{\theta}^{(t)})^\top J(\theta^{(t)})\boldsymbol{Q}^{(t)}\right\|_2\right) \left\|g(\theta^{(t)})\right\|_2 \\ &\leq \left(1 + \eta \left\|J(\tilde{\theta}^{(t)})\right\|_F \left\|J(\theta^{(t)})\right\|_F\right) \left\|g(\theta^{(t)})\right\|_2 \\ &\leq (1 + \eta^* M^2) \left\|g(\theta^{(t)})\right\|_2 \leq B^{t+1} R_0\end{aligned} \tag{56}$$

Therefore, (48) and (49) are true for all $t \leq t_\epsilon$, which implies that $\left\|\sqrt{\boldsymbol{Q}} g(\theta^{(t_\epsilon)})\right\|_2 \leq \left\|g(\theta^{(t_\epsilon)})\right\|_2 \leq B^{t_\epsilon} R_0$, so (50) is true for $t = t_\epsilon$. And (51) is obviously true for $t = t_\epsilon$. Now, let us prove (ii) by induction. Note that when $t \geq t_\epsilon$, we have the alternative update rule (47). If (50) and (51) are true for t, then for $t + 1$, there is

$$\begin{aligned}\left\|\theta^{(t+1)} - \theta^{(t)}\right\|_2 &\leq \eta \left\|J(\theta^{(t)})\sqrt{\boldsymbol{Q}_{3\epsilon}^{(t)}}\right\|_2 \left\|\sqrt{\boldsymbol{Q}} g(\theta^{(t)})\right\|_2 \leq \eta \left\|J(\theta^{(t)})\sqrt{\boldsymbol{Q}_{3\epsilon}^{(t)}}\right\|_F \left\|\sqrt{\boldsymbol{Q}} g(\theta^{(t)})\right\|_2 \\ &\leq \eta\sqrt{1+3\epsilon} \left\|J(\theta^{(t)})\right\|_F \left\|\sqrt{\boldsymbol{Q}} g(\theta^{(t)})\right\|_2 \leq M\eta\sqrt{1+3\epsilon} \left(1 - \frac{\eta q^* \lambda^{\min}}{3}\right)^{t-t_\epsilon} B^{t_\epsilon} R_0\end{aligned} \tag{57}$$

So (51) is true for $t+1$. And we also have

$$\begin{aligned}\left\|\sqrt{\boldsymbol{Q}}g(\theta^{(t+1)})\right\|_2 &= \left\|\sqrt{\boldsymbol{Q}}g(\theta^{(t+1)}) - \sqrt{\boldsymbol{Q}}g(\theta^{(t)}) + \sqrt{\boldsymbol{Q}}g(\theta^{(t)})\right\|_2 \\ &= \left\|\sqrt{\boldsymbol{Q}}J(\tilde{\theta}^{(t)})^\top(\theta^{(t+1)} - \theta^{(t)}) + \sqrt{\boldsymbol{Q}}g(\theta^{(t)})\right\|_2 \\ &= \left\|-\eta\sqrt{\boldsymbol{Q}}J(\tilde{\theta}^{(t)})^\top J(\theta^{(t)})\boldsymbol{Q}^{(t)}g(\theta^{(t)}) + \sqrt{\boldsymbol{Q}}g(\theta^{(t)})\right\|_2 \\ &\leq \left\|\boldsymbol{I} - \eta\sqrt{\boldsymbol{Q}}J(\tilde{\theta}^{(t)})^\top J(\theta^{(t)})\sqrt{\boldsymbol{Q}_{3\epsilon}^{(t)}}\right\|_2 \left\|\sqrt{\boldsymbol{Q}}g(\theta^{(t)})\right\|_2 \\ &\leq \left\|\boldsymbol{I} - \eta\sqrt{\boldsymbol{Q}}J(\tilde{\theta}^{(t)})^\top J(\theta^{(t)})\sqrt{\boldsymbol{Q}_{3\epsilon}^{(t)}}\right\|_2 \left(1 - \frac{\eta q^* \lambda^{\min}}{3}\right)^t R_0\end{aligned} \tag{58}$$

where $\tilde{\theta}^{(t)}$ is some linear interpolation between $\theta^{(t)}$ and $\theta^{(t+1)}$. Now we prove that

$$\left\|\boldsymbol{I} - \eta\sqrt{\boldsymbol{Q}}J(\tilde{\theta}^{(t)})^\top J(\theta^{(t)})\sqrt{\boldsymbol{Q}_{3\epsilon}^{(t)}}\right\|_2 \leq 1 - \frac{\eta q^* \lambda^{\min}}{3} \tag{59}$$

For any unit vector $\boldsymbol{v} \in \mathbb{R}^n$, we have

$$\boldsymbol{v}^\top(\boldsymbol{I} - \eta\sqrt{\boldsymbol{Q}}\Theta\sqrt{\boldsymbol{Q}})\boldsymbol{v} = 1 - \eta\boldsymbol{v}^\top\sqrt{\boldsymbol{Q}}\Theta\sqrt{\boldsymbol{Q}}\boldsymbol{v} \tag{60}$$

$\left\|\sqrt{\boldsymbol{Q}}\boldsymbol{v}\right\|_2 \in [\sqrt{q^*}, 1]$, so for any $\eta \leq \eta^*$, $\boldsymbol{v}^\top(\boldsymbol{I} - \eta\sqrt{\boldsymbol{Q}}\Theta\sqrt{\boldsymbol{Q}})\boldsymbol{v} \in [0, 1 - \eta\lambda^{\min}q^*]$, which implies that $\left\|\boldsymbol{I} - \eta\sqrt{\boldsymbol{Q}}\Theta\sqrt{\boldsymbol{Q}}\right\|_2 \leq 1 - \eta\lambda^{\min}q^*$. Thus,

$$\begin{aligned}&\left\|\boldsymbol{I} - \eta\sqrt{\boldsymbol{Q}}J(\tilde{\theta}^{(t)})^\top J(\theta^{(t)})\sqrt{\boldsymbol{Q}}\right\|_2 \\ \leq&\left\|\boldsymbol{I} - \eta\sqrt{\boldsymbol{Q}}\Theta\sqrt{\boldsymbol{Q}}\right\|_2 + \eta\left\|\sqrt{\boldsymbol{Q}}(\Theta - \Theta^{(0)})\sqrt{\boldsymbol{Q}}\right\|_2 + \eta\left\|\sqrt{\boldsymbol{Q}}(J(\theta^{(0)})^\top J(\theta^{(0)}) - J(\tilde{\theta}^{(t)})^\top J(\theta^{(t)}))\sqrt{\boldsymbol{Q}}\right\|_2 \\ \leq&1 - \eta\lambda^{\min}q^* + \eta\left\|\sqrt{\boldsymbol{Q}}(\Theta - \Theta^{(0)})\sqrt{\boldsymbol{Q}}\right\|_F + \eta\left\|\sqrt{\boldsymbol{Q}}(J(\theta^{(0)})^\top J(\theta^{(0)}) - J(\tilde{\theta}^{(t)})^\top J(\theta^{(t)}))\sqrt{\boldsymbol{Q}}\right\|_F \\ \leq&1 - \eta\lambda^{\min}q^* + \eta\left\|\Theta - \Theta^{(0)}\right\|_F + \eta\left\|J(\theta^{(0)})^\top J(\theta^{(0)}) - J(\tilde{\theta}^{(t)})^\top J(\theta^{(t)})\right\|_F \\ \leq&1 - \eta\lambda^{\min}q^* + \frac{\eta q^*\lambda^{\min}}{3} + \frac{\eta M^2}{\sqrt[4]{\tilde{d}}}\left(\left\|\theta^{(t)} - \theta^{(0)}\right\|_2 + \left\|\tilde{\theta}^{(t)} - \theta^{(0)}\right\|_2\right) \leq 1 - \frac{\eta q^*\lambda^{\min}}{2}\end{aligned} \tag{61}$$

for all $\tilde{d} \geq \max\left\{D_1, D_2, \left(\frac{12M^2C_0}{q^*\lambda^{\min}}\right)^4\right\}$, which implies that

$$\begin{aligned}&\left\|\boldsymbol{I} - \eta\sqrt{\boldsymbol{Q}}J(\tilde{\theta}^{(t)})^\top J(\theta^{(t)})\sqrt{\boldsymbol{Q}_{3\epsilon}^{(t)}}\right\|_2 \\ \leq&1 - \frac{\eta q^*\lambda^{\min}}{2} + \left\|\eta\sqrt{\boldsymbol{Q}}J(\tilde{\theta}^{(t)})^\top J(\theta^{(t)})\left(\sqrt{\boldsymbol{Q}_{3\epsilon}^{(t)}} - \sqrt{\boldsymbol{Q}}\right)\right\|_2 \\ \leq&1 - \frac{\eta q^*\lambda^{\min}}{2} + \eta M^2\sqrt{3\epsilon} \leq 1 - \frac{\eta q^*\lambda^{\min}}{3} \qquad \text{(due to (37))}\end{aligned} \tag{62}$$

for all $\epsilon \leq \epsilon_0$. Thus, (50) is also true for $t+1$. In conclusion, (50) and (51) are true with probability at least $(1-\delta)$ for all $\tilde{d} \geq \tilde{D} = \max\left\{D_1, D_2, \left(\frac{12M^2C_0}{q^*\lambda^{\min}}\right)^4\right\}$. □

Returning back to the proof of Lemma 5. Choose and fix an ϵ such that $\epsilon < \min\{\epsilon_0, \frac{1}{3}\left(\frac{q^*\lambda^{\min}}{3\lambda^{\max}+q^*\lambda^{\min}}\right)^2\}$, where ϵ_0 is defined by Theorem 12. Then, t_ϵ is also fixed. There exists $\tilde{D} \geq 0$ such that for any $\tilde{d} \geq \tilde{D}$, with probability at least $(1-\delta)$, Theorem 12 and Lemma 13 are true and

$$\left\|\Theta - \Theta^{(0)}\right\|_F \leq \frac{q^*\lambda^{\min}}{3} \tag{63}$$

which immediately implies that

$$\left\|\Theta^{(0)}\right\|_2 \leq \|\Theta\|_2 + \left\|\Theta - \Theta^{(0)}\right\|_F \leq \lambda^{\max} + \frac{q^*\lambda^{\min}}{3} \tag{64}$$

We still denote $B = 1 + \eta^* M^2$ and $C_0 = \frac{MB^{t_\epsilon}R_0}{B-1} + \frac{3\sqrt{1+3\epsilon}MB^{t_\epsilon}R_0}{q^*\lambda^{\min}}$. Theorem 12 ensures that for all t, $\theta^{(t)} \in B(\theta^{(0)}, C_0)$. Then we have

$$\begin{aligned}\left\|\boldsymbol{I} - \eta\sqrt{\boldsymbol{Q}}\Theta^{(0)}\sqrt{\boldsymbol{Q}}\right\|_2 &\leq \left\|\boldsymbol{I} - \eta\sqrt{\boldsymbol{Q}}\Theta\sqrt{\boldsymbol{Q}}\right\|_2 + \eta\left\|\sqrt{\boldsymbol{Q}}(\Theta - \Theta^{(0)})\sqrt{\boldsymbol{Q}}\right\|_2 \\ &\leq 1 - \eta\lambda^{\min}q^* + \frac{\eta q^*\lambda^{\min}}{3} = 1 - \frac{2\eta q^*\lambda^{\min}}{3}\end{aligned} \tag{65}$$

so it follows that

$$\begin{aligned}\left\|\boldsymbol{I} - \eta\sqrt{\boldsymbol{Q}}\Theta^{(0)}\sqrt{\boldsymbol{Q}_{3\epsilon}^{(t)}}\right\|_2 &\leq \left\|\boldsymbol{I} - \eta\sqrt{\boldsymbol{Q}}\Theta^{(0)}\sqrt{\boldsymbol{Q}}\right\|_2 + \left\|\eta\sqrt{\boldsymbol{Q}}\Theta^{(0)}\left(\sqrt{\boldsymbol{Q}_{3\epsilon}^{(t)}} - \sqrt{\boldsymbol{Q}}\right)\right\|_2 \\ &\leq 1 - \frac{2\eta q^*\lambda^{\min}}{3} + \eta(\lambda^{\max} + \frac{q^*\lambda^{\min}}{3})\sqrt{3\epsilon}\end{aligned} \tag{66}$$

Thus, for all $\epsilon < \frac{1}{3}\left(\frac{q^*\lambda^{\min}}{3\lambda^{\max}+q^*\lambda^{\min}}\right)^2$, there is

$$\left\|\boldsymbol{I} - \eta\sqrt{\boldsymbol{Q}}\Theta^{(0)}\sqrt{\boldsymbol{Q}_{3\epsilon}^{(t)}}\right\|_2 \leq 1 - \frac{\eta q^*\lambda^{\min}}{3} \tag{67}$$

The update rule of the GRW for the linearized neural network is:

$$\theta_{\text{lin}}^{(t+1)} = \theta_{\text{lin}}^{(t)} - \eta J(\theta^{(0)})\boldsymbol{Q}^{(t)}g_{\text{lin}}(\theta^{(t)}) \tag{68}$$

where we use the subscript "lin" to denote the linearized neural network, and with a slight abuse of notion denote $g_{\text{lin}}(\theta^{(t)}) = g(\theta_{\text{lin}}^{(t)})$.

First, let us consider the training data $\boldsymbol{X}$. Denote $\Delta_t = g_{\text{lin}}(\theta^{(t)}) - g(\theta^{(t)})$. We have

$$\begin{cases} g_{\text{lin}}(\theta^{(t+1)}) - g_{\text{lin}}(\theta^{(t)}) = -\eta J(\theta^{(0)})^\top J(\theta^{(0)})\boldsymbol{Q}^{(t)}g_{\text{lin}}(\theta^{(t)}) \\ g(\theta^{(t+1)}) - g(\theta^{(t)}) = -\eta J(\tilde{\theta}^{(t)})^\top J(\theta^{(t)})\boldsymbol{Q}^{(t)}g(\theta^{(t)}) \end{cases} \tag{69}$$

where $\tilde{\theta}^{(t)}$ is some linear interpolation between $\theta^{(t)}$ and $\theta^{(t+1)}$. Thus,

$$\begin{aligned}\Delta_{t+1} - \Delta_t =& \eta\left[J(\tilde{\theta}^{(t)})^\top J(\theta^{(t)}) - J(\theta^{(0)})^\top J(\theta^{(0)})\right]\boldsymbol{Q}^{(t)}g(\theta^{(t)}) \\ &- \eta J(\theta^{(0)})^\top J(\theta^{(0)})\boldsymbol{Q}^{(t)}\Delta_t\end{aligned} \tag{70}$$

By Lemma 13, we have

$$\begin{aligned}&\left\|J(\tilde{\theta}^{(t)})^\top J(\theta^{(t)}) - J(\theta^{(0)})^\top J(\theta^{(0)})\right\|_F \\ \leq& \left\|\left(J(\tilde{\theta}^{(t)}) - J(\theta^{(0)})\right)^\top J(\theta^{(t)})\right\|_F + \left\|J(\theta^{(0)})^\top\left(J(\theta^{(t)}) - J(\theta^{(0)})\right)\right\|_F \\ \leq& 2M^2C_0\tilde{d}^{-1/4}\end{aligned} \tag{71}$$

which implies that for all $t < t_\epsilon$,

$$\begin{aligned}\|\Delta_{t+1}\|_2 &\leq \left\|\left[\boldsymbol{I} - \eta J(\theta^{(0)})^\top J(\theta^{(0)})\boldsymbol{Q}^{(t)}\right]\Delta_t\right\|_2 + \left\|\eta\left[J(\tilde{\theta}^{(t)})^\top J(\theta^{(t)}) - J(\theta^{(0)})^\top J(\theta^{(0)})\right]\boldsymbol{Q}^{(t)}g(\theta^{(t)})\right\|_2 \\ &\leq \left\|\boldsymbol{I} - \eta J(\theta^{(0)})^\top J(\theta^{(0)})\boldsymbol{Q}^{(t)}\right\|_F\|\Delta_t\|_2 + \eta\left\|J(\tilde{\theta}^{(t)})^\top J(\theta^{(t)}) - J(\theta^{(0)})^\top J(\theta^{(0)})\right\|_F\left\|g(\theta^{(t)})\right\|_2 \\ &\leq (1 + \eta M^2)\|\Delta_t\|_2 + 2\eta M^2C_0B^tR_0\tilde{d}^{-1/4} \\ &\leq B\|\Delta_t\|_2 + 2\eta M^2C_0B^tR_0\tilde{d}^{-1/4}\end{aligned} \tag{72}$$

Therefore, we have

$$B^{-(t+1)} \left\|\Delta_{t+1}\right\|_2 \leq B^{-t} \left\|\Delta_t\right\|_2 + 2\eta M^2 C_0 B^{-1} R_0 \tilde{d}^{-1/4} \tag{73}$$

Since $\Delta_0 = 0$, it follows that for all $t \leq t_\epsilon$,

$$\left\|\Delta_t\right\|_2 \leq 2t\eta M^2 C_0 B^{t-1} R_0 \tilde{d}^{-1/4} \tag{74}$$

and particularly we have

$$\left\|\sqrt{\boldsymbol{Q}}\Delta_{t_\epsilon}\right\|_2 \leq \left\|\Delta_{t_\epsilon}\right\|_2 \leq 2t_\epsilon \eta M^2 C_0 B^{t_\epsilon - 1} R_0 \tilde{d}^{-1/4} \tag{75}$$

For $t \geq t_\epsilon$, we have the alternative update rule (47). Thus,

$$\begin{aligned}\sqrt{\boldsymbol{Q}}\Delta_{t+1} - \sqrt{\boldsymbol{Q}}\Delta_t =& \eta\sqrt{\boldsymbol{Q}}\left[J(\tilde{\theta}^{(t)})^\top J(\theta^{(t)}) - J(\theta^{(0)})^\top J(\theta^{(0)})\right]\sqrt{\boldsymbol{Q}_{3\epsilon}^{(t)}}\left[\sqrt{\boldsymbol{Q}}g(\theta^{(t)})\right] \\ & - \eta\sqrt{\boldsymbol{Q}}J(\theta^{(0)})^\top J(\theta^{(0)})\sqrt{\boldsymbol{Q}_{3\epsilon}^{(t)}}\left[\sqrt{\boldsymbol{Q}}\Delta_t\right]\end{aligned} \tag{76}$$

Let $\boldsymbol{A} = \boldsymbol{I} - \eta\sqrt{\boldsymbol{Q}}J(\theta^{(0)})^\top J(\theta^{(0)})\sqrt{\boldsymbol{Q}_{3\epsilon}^{(t)}} = \boldsymbol{I} - \eta\sqrt{\boldsymbol{Q}}\Theta^{(0)}\sqrt{\boldsymbol{Q}_{3\epsilon}^{(t)}}$. Then, we have

$$\sqrt{\boldsymbol{Q}}\Delta_{t+1} = \boldsymbol{A}\sqrt{\boldsymbol{Q}}\Delta_t + \eta\sqrt{\boldsymbol{Q}}\left[J(\tilde{\theta}^{(t)})^\top J(\theta^{(t)}) - J(\theta^{(0)})^\top J(\theta^{(0)})\right]\sqrt{\boldsymbol{Q}_{3\epsilon}^{(t)}}\left(\sqrt{\boldsymbol{Q}}g(\theta^{(t)})\right) \tag{77}$$

Let $\gamma = 1 - \frac{\eta q^* \lambda^{\min}}{3} < 1$. Combining with Theorem 12 and (67), the above leads to

$$\begin{aligned}\left\|\sqrt{\boldsymbol{Q}}\Delta_{t+1}\right\|_2 &\leq \left\|\boldsymbol{A}\right\|_2 \left\|\sqrt{\boldsymbol{Q}}\Delta_t\right\|_2 + \eta\left\|\sqrt{\boldsymbol{Q}}\left[J(\tilde{\theta}^{(t)})^\top J(\theta^{(t)}) - J(\theta^{(0)})^\top J(\theta^{(0)})\right]\sqrt{\boldsymbol{Q}_{3\epsilon}^{(t)}}\right\|_2 \left\|\sqrt{\boldsymbol{Q}}g(\theta^{(t)})\right\|_2 \\ &\leq \gamma\left\|\sqrt{\boldsymbol{Q}}\Delta_t\right\|_2 + \eta\left\|J(\tilde{\theta}^{(t)})^\top J(\theta^{(t)}) - J(\theta^{(0)})^\top J(\theta^{(0)})\right\|_F \sqrt{1+3\epsilon}\gamma^{t-t_\epsilon}B^{t_\epsilon}R_0 \\ &\leq \gamma\left\|\sqrt{\boldsymbol{Q}}\Delta_t\right\|_2 + 2\eta M^2 C_0\sqrt{1+3\epsilon}\gamma^{t-t_\epsilon}B^{t_\epsilon}R_0\tilde{d}^{-1/4}\end{aligned} \tag{78}$$

This implies that

$$\gamma^{-(t+1)}\left\|\sqrt{\boldsymbol{Q}}\Delta_{t+1}\right\|_2 \leq \gamma^{-t}\left\|\sqrt{\boldsymbol{Q}}\Delta_t\right\|_2 + 2\eta M^2 C_0\sqrt{1+3\epsilon}\gamma^{-1-t_\epsilon}B^{t_\epsilon}R_0\tilde{d}^{-1/4} \tag{79}$$

Combining with (75), it implies that for all $t \geq t_\epsilon$,

$$\left\|\sqrt{\boldsymbol{Q}}\Delta_t\right\|_2 \leq 2\gamma^{t-t_\epsilon}\eta M^2 C_0 B^{t_\epsilon} R_0\left[t_\epsilon B^{-1} + \sqrt{1+3\epsilon}\gamma^{-1}(t-t_\epsilon)\right]\tilde{d}^{-1/4} \tag{80}$$

Next, we consider an arbitrary test point $\boldsymbol{x}$ such that $\left\|\boldsymbol{x}\right\|_2 \leq 1$. Denote $\delta_t = f_{\text{lin}}^{(t)}(\boldsymbol{x}) - f^{(t)}(\boldsymbol{x})$. Then we have

$$\begin{cases} f_{\text{lin}}^{(t+1)}(\boldsymbol{x}) - f_{\text{lin}}^{(t)}(\boldsymbol{x}) = -\eta\nabla_\theta f(\boldsymbol{x};\theta^{(0)})^\top J(\theta^{(0)})\boldsymbol{Q}^{(t)}g_{\text{lin}}(\theta^{(t)}) \\ f^{(t+1)}(\boldsymbol{x}) - f^{(t)}(\boldsymbol{x}) = -\eta\nabla_\theta f(\boldsymbol{x};\tilde{\theta}^{(t)})^\top J(\theta^{(t)})\boldsymbol{Q}^{(t)}g(\theta^{(t)}) \end{cases} \tag{81}$$

which yields

$$\begin{aligned}\delta_{t+1} - \delta_t =& \eta\left[\nabla_\theta f(\boldsymbol{x};\tilde{\theta}^{(t)})^\top J(\theta^{(t)}) - \nabla_\theta f(\boldsymbol{x};\theta^{(0)})^\top J(\theta^{(0)})\right]\boldsymbol{Q}^{(t)}g(\theta^{(t)}) \\ & - \eta\nabla_\theta f(\boldsymbol{x};\theta^{(0)})^\top J(\theta^{(0)})\boldsymbol{Q}^{(t)}\Delta_t\end{aligned} \tag{82}$$

For $t \leq t_\epsilon$, we have

$$\begin{aligned}
\|\delta_t\|_2 \leq & \eta \sum_{s=0}^{t-1} \left\| \left[\nabla_\theta f(\boldsymbol{x}; \tilde{\theta}^{(s)})^\top J(\theta^{(s)}) - \nabla_\theta f(\boldsymbol{x}; \theta^{(0)})^\top J(\theta^{(0)}) \right] \boldsymbol{Q}^{(s)} \right\|_2 \left\| g(\theta^{(s)}) \right\|_2 \\
& + \eta \sum_{s=0}^{t-1} \left\| \nabla_\theta f(\boldsymbol{x}; \theta^{(0)})^\top J(\theta^{(0)}) \boldsymbol{Q}^{(s)} \right\|_2 \|\Delta_s\|_2 \\
\leq & \eta \sum_{s=0}^{t-1} \left\| \nabla_\theta f(\boldsymbol{x}; \tilde{\theta}^{(s)})^\top J(\theta^{(s)}) - \nabla_\theta f(\boldsymbol{x}; \theta^{(0)})^\top J(\theta^{(0)}) \right\|_F \left\| g(\theta^{(s)}) \right\|_2 \\
& + \eta \sum_{s=0}^{t-1} \left\| \nabla_\theta f(\boldsymbol{x}; \theta^{(0)}) \right\|_2 \left\| J(\theta^{(0)}) \right\|_F \|\Delta_s\|_2 \\
\leq & 2\eta M^2 C_0 \tilde{d}^{-1/4} \sum_{s=0}^{t-1} B^s R_0 + \eta M^2 \sum_{s=0}^{t-1} (2s\eta M^2 C_0 B^{s-1} R_0 \tilde{d}^{-1/4})
\end{aligned} \tag{83}$$

So we can see that there exists a constant C_1 such that $\|\delta_{t_\epsilon}\|_2 \leq C_1 \tilde{d}^{-1/4}$. Then, for $t > t_\epsilon$, we have

$$\begin{aligned}
\|\delta_t\|_2 - \|\delta_{t_\epsilon}\|_2 \leq & \eta \sum_{s=t_\epsilon}^{t-1} \left\| \left[\nabla_\theta f(\boldsymbol{x}; \tilde{\theta}^{(s)})^\top J(\theta^{(s)}) - \nabla_\theta f(\boldsymbol{x}; \theta^{(0)})^\top J(\theta^{(0)}) \right] \sqrt{\boldsymbol{Q}_{3\epsilon}^{(s)}} \right\|_2 \left\| \sqrt{\boldsymbol{Q}} g(\theta^{(s)}) \right\|_2 \\
& + \eta \sum_{s=t_\epsilon}^{t-1} \left\| \nabla_\theta f(\boldsymbol{x}; \theta^{(0)})^\top J(\theta^{(0)}) \sqrt{\boldsymbol{Q}_{3\epsilon}^{(s)}} \right\|_2 \left\| \sqrt{\boldsymbol{Q}} \Delta_s \right\|_2 \\
\leq & 2\eta M^2 C_0 \tilde{d}^{-1/4} \sqrt{1+3\epsilon} \sum_{s=t_\epsilon}^{t-1} \gamma^{s-t_\epsilon} B^{t_\epsilon} R_0 \\
& + \eta M^2 \sqrt{1+3\epsilon} \sum_{s=t_\epsilon}^{t-1} \left(2\gamma^{s-t_\epsilon} \eta M^2 C_0 B^{t_\epsilon} R_0 \left[t_\epsilon B^{-1} + \sqrt{1+3\epsilon} \gamma^{-1} (s - t_\epsilon) \right] \tilde{d}^{-1/4} \right)
\end{aligned} \tag{84}$$

Note that $\sum_{t=0}^{\infty} t\gamma^t$ is finite as long as $\gamma \in (0,1)$. Therefore, there is a constant C such that for any t, $\|\delta_t\|_2 \leq C\tilde{d}^{-1/4}$ with probability at least $(1-\delta)$ for any $\tilde{d} \geq \tilde{D}$. □

D.3.3 Proof of Lemma 13

We will use the following theorem regarding the eigenvalues of random Gaussian matrices:

Theorem 14 (Corollary 5.35 in [Ver10]). *If $\boldsymbol{A} \in \mathbb{R}^{p \times q}$ is a random matrix whose entries are independent standard normal random variables, then for every $t \geq 0$, with probability at least $1 - 2\exp(-t^2/2)$,*

$$\sqrt{p} - \sqrt{q} - t \leq \lambda^{\min}(\boldsymbol{A}) \leq \lambda^{\max}(\boldsymbol{A}) \leq \sqrt{p} + \sqrt{q} + t \tag{85}$$

By this theorem, and also note that W^L is a vector, we can see that for any δ, there exist $\tilde{D} > 0$ and $M_1 > 0$ such that if $\tilde{d} \geq \tilde{D}$, then with probability at least $(1-\delta)$, for all $\theta \in B(\theta^{(0)}, C_0)$, we have

$$\left\|W^l\right\|_2 \leq 3\sqrt{\tilde{d}} \quad (\forall 0 \leq l \leq L-1) \qquad \text{and} \qquad \left\|W^L\right\|_2 \leq C_0 \leq 3\sqrt[4]{\tilde{d}} \tag{86}$$

as well as

$$\left\|\beta \boldsymbol{b}^l\right\|_2 \leq M_1 \sqrt{\tilde{d}} \qquad (\forall l = 0, \cdots, L) \tag{87}$$

Now we assume that (86) and (87) are true. Then, for any $\boldsymbol{x}$ such that $\|\boldsymbol{x}\|_2 \leq 1$,

$$\begin{aligned}
\left\|\boldsymbol{h}^1\right\|_2 &= \left\| \frac{1}{\sqrt{d_0}} W^0 \boldsymbol{x} + \beta \boldsymbol{b}^0 \right\|_2 \leq \frac{1}{\sqrt{d_0}} \left\|W^0\right\|_2 \|\boldsymbol{x}\|_2 + \left\|\beta \boldsymbol{b}^0\right\|_2 \leq (\frac{3}{\sqrt{d_0}} + M_1)\sqrt{\tilde{d}} \\
\left\|\boldsymbol{h}^{l+1}\right\|_2 &= \left\| \frac{1}{\sqrt{\tilde{d}}} W^l \boldsymbol{x}^l + \beta \boldsymbol{b}^l \right\|_2 \leq \frac{1}{\sqrt{\tilde{d}}} \left\|W^l\right\|_2 \left\|\boldsymbol{x}^l\right\|_2 + \left\|\beta \boldsymbol{b}^l\right\|_2 \qquad (\forall l \geq 1) \\
\left\|\boldsymbol{x}^l\right\|_2 &= \left\|\sigma(\boldsymbol{h}^l) - \sigma(\boldsymbol{0}^l) + \sigma(\boldsymbol{0}^l)\right\|_2 \leq L_0 \left\|\boldsymbol{h}^l\right\|_2 + \sigma(0)\sqrt{\tilde{d}} \qquad (\forall l \geq 1)
\end{aligned} \tag{88}$$

where L_0 is the Lipschitz constant of σ and $\sigma(\mathbf{0}^l) = (\sigma(0), \cdots, \sigma(0)) \in \mathbb{R}^{d_l}$. By induction, there exists an $M_2 > 0$ such that $\left\|\boldsymbol{x}^l\right\|_2 \leq M_2\sqrt{\tilde{d}}$ and $\left\|\boldsymbol{h}^l\right\|_2 \leq M_2\sqrt{\tilde{d}}$ for all $l = 1, \cdots, L$.

Denote $\boldsymbol{\alpha}^l = \nabla_{\boldsymbol{h}^l} f(\boldsymbol{x}) = \nabla_{\boldsymbol{h}^l} \boldsymbol{h}^{L+1}$. For all $l = 1, \cdots, L$, we have $\boldsymbol{\alpha}^l = \mathrm{diag}(\dot{\sigma}(\boldsymbol{h}^l))\frac{W^{l\top}}{\sqrt{\tilde{d}}}\boldsymbol{\alpha}^{l+1}$ where $\dot{\sigma}(x) \leq L_0$ for all $x \in \mathbb{R}$ since σ is L_0-Lipschitz, $\boldsymbol{\alpha}^{L+1} = 1$ and $\left\|\boldsymbol{\alpha}^L\right\|_2 = \left\|\mathrm{diag}(\dot{\sigma}(\boldsymbol{h}^L))\frac{W^{L\top}}{\sqrt{\tilde{d}}}\right\|_2 \leq \frac{3}{\sqrt[4]{\tilde{d}}}L_0$. Then, we can easily prove by induction that there exists an $M_3 > 1$ such that $\left\|\boldsymbol{\alpha}^l\right\|_2 \leq M_3/\sqrt[4]{\tilde{d}}$ for all $l = 1, \cdots, L$ (note that this is not true for $L+1$ because $\boldsymbol{\alpha}^{L+1} = 1$).

For $l = 0$, $\nabla_{W^0} f(\boldsymbol{x}) = \frac{1}{\sqrt{d_0}}\boldsymbol{x}^0\boldsymbol{\alpha}^{1\top}$, so $\|\nabla_{W^l} f(\boldsymbol{x})\|_2 \leq \frac{1}{\sqrt{d_0}}\left\|\boldsymbol{x}^0\right\|_2\left\|\boldsymbol{\alpha}^1\right\|_2 \leq \frac{1}{\sqrt{d_0}}M_3/\sqrt[4]{\tilde{d}}$. And for any $l = 1, \cdots, L$, $\nabla_{W^l} f(\boldsymbol{x}) = \frac{1}{\sqrt{\tilde{d}}}\boldsymbol{x}^l\boldsymbol{\alpha}^{l+1}$, so $\|\nabla_{W^l} f(\boldsymbol{x})\|_2 \leq \frac{1}{\sqrt{\tilde{d}}}\left\|\boldsymbol{x}^l\right\|_2\left\|\boldsymbol{\alpha}^{l+1}\right\|_2 \leq M_2M_3$. (Note that if $M_3 > 1$, then $\left\|\boldsymbol{\alpha}^{L+1}\right\|_2 \leq M_3$; and since $\tilde{d} \geq 1$, there is $\left\|\boldsymbol{\alpha}^l\right\|_2 \leq M_3$ for $l \leq L$.) Moreover, for $l = 0, \cdots, L$, $\nabla_{\boldsymbol{b}^l} f(\boldsymbol{x}) = \beta\boldsymbol{\alpha}^{l+1}$, so $\|\nabla_{\boldsymbol{b}^l} f(\boldsymbol{x})\|_2 \leq \beta M_3$. Thus, if (86) and (87) are true, then there exists an $M_4 > 0$, such that $\|\nabla_\theta f(\boldsymbol{x})\|_2 \leq M_4/\sqrt{n}$. And since $\|\boldsymbol{x}_i\|_2 \leq 1$ for all i, so $\|J(\theta)\|_F \leq M_4$.

Next, we consider the difference in $\nabla_\theta f(\boldsymbol{x})$ between θ and $\tilde{\theta}$. Let $\tilde{f}, \tilde{W}, \tilde{\boldsymbol{b}}, \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{h}}, \tilde{\boldsymbol{\alpha}}$ be the function and the values corresponding to $\tilde{\theta}$. There is

$$
\begin{aligned}
\left\|\boldsymbol{h}^1 - \tilde{\boldsymbol{h}}^1\right\|_2 &= \left\|\frac{1}{\sqrt{d_0}}(W^0 - \tilde{W}^0)\boldsymbol{x} + \beta(\boldsymbol{b}^0 - \tilde{\boldsymbol{b}}^0)\right\|_2 \\
&\leq \frac{1}{\sqrt{d_0}}\left\|W^0 - \tilde{W}^0\right\|_2\|\boldsymbol{x}\|_2 + \beta\left\|\boldsymbol{b}^0 - \tilde{\boldsymbol{b}}^0\right\|_2 \leq \left(\frac{1}{\sqrt{d_0}} + \beta\right)\left\|\theta - \tilde{\theta}\right\|_2 \\
\left\|\boldsymbol{h}^{l+1} - \tilde{\boldsymbol{h}}^{l+1}\right\|_2 &= \left\|\frac{1}{\sqrt{\tilde{d}}}W^l(\boldsymbol{x}^l - \tilde{\boldsymbol{x}}^l) + \frac{1}{\sqrt{\tilde{d}}}(W^l - \tilde{W}^l)\tilde{\boldsymbol{x}}^l + \beta(\boldsymbol{b}^l - \tilde{\boldsymbol{b}}^l)\right\|_2 \\
&\leq \frac{1}{\sqrt{\tilde{d}}}\left\|W^l\right\|_2\left\|\boldsymbol{x}^l - \tilde{\boldsymbol{x}}^l\right\|_2 + \frac{1}{\sqrt{\tilde{d}}}\left\|W^l - \tilde{W}^l\right\|_2\left\|\tilde{\boldsymbol{x}}^l\right\|_2 + \beta\left\|\boldsymbol{b}^l - \tilde{\boldsymbol{b}}^l\right\|_2 \\
&\leq 3\left\|\boldsymbol{x}^l - \tilde{\boldsymbol{x}}^l\right\|_2 + (M_2 + \beta)\left\|\theta - \tilde{\theta}\right\|_2 \qquad (\forall l \geq 1) \\
\left\|\boldsymbol{x}^l - \tilde{\boldsymbol{x}}^l\right\|_2 &= \left\|\sigma(\boldsymbol{h}^l) - \sigma(\tilde{\boldsymbol{h}}^l)\right\|_2 \leq L_0\left\|\boldsymbol{h}^l - \tilde{\boldsymbol{h}}^l\right\|_2 \qquad (\forall l \geq 1)
\end{aligned}
\tag{89}
$$

By induction, there exists an $M_5 > 0$ such that $\left\|\boldsymbol{x}^l - \tilde{\boldsymbol{x}}^l\right\|_2 \leq M_5\left\|\theta - \tilde{\theta}\right\|_2$ for all l.

For $\boldsymbol{\alpha}^l$, we have $\boldsymbol{\alpha}^{L+1} = \tilde{\boldsymbol{\alpha}}^{L+1} = 1$, and for all $l \geq 1$,

$$
\begin{aligned}
\left\|\boldsymbol{\alpha}^l - \tilde{\boldsymbol{\alpha}}^l\right\|_2 &= \left\|\mathrm{diag}(\dot{\sigma}(\boldsymbol{h}^l))\frac{W^{l\top}}{\sqrt{\tilde{d}}}\boldsymbol{\alpha}^{l+1} - \mathrm{diag}(\dot{\sigma}(\tilde{\boldsymbol{h}}^l))\frac{\tilde{W}^{l\top}}{\sqrt{\tilde{d}}}\tilde{\boldsymbol{\alpha}}^{l+1}\right\|_2 \\
&\leq \left\|\mathrm{diag}(\dot{\sigma}(\boldsymbol{h}^l))\frac{W^{l\top}}{\sqrt{\tilde{d}}}(\boldsymbol{\alpha}^{l+1} - \tilde{\boldsymbol{\alpha}}^{l+1})\right\|_2 + \left\|\mathrm{diag}(\dot{\sigma}(\boldsymbol{h}^l))\frac{(W^l - \tilde{W}^l)^\top}{\sqrt{\tilde{d}}}\tilde{\boldsymbol{\alpha}}^{l+1}\right\|_2 \\
&\qquad + \left\|\mathrm{diag}((\dot{\sigma}(\boldsymbol{h}^l) - \dot{\sigma}(\tilde{\boldsymbol{h}}^l)))\frac{\tilde{W}^{l\top}}{\sqrt{\tilde{d}}}\tilde{\boldsymbol{\alpha}}^{l+1}\right\|_2 \\
&\leq 3L_0\left\|\boldsymbol{\alpha}^{l+1} - \tilde{\boldsymbol{\alpha}}^{l+1}\right\|_2 + \left(M_3L_0\tilde{d}^{-1/2} + 3M_3M_5L_1\tilde{d}^{-1/4}\right)\left\|\theta - \tilde{\theta}\right\|_2
\end{aligned}
\tag{90}
$$

where L_1 is the Lipschitz constant of $\dot{\sigma}$. Particularly, for $l = L$, though $\tilde{\boldsymbol{\alpha}}^{L+1} = 1$, since $\left\|\tilde{W}^L\right\|_2 \leq 3\tilde{d}^{1/4}$, (90) is still true. By induction, there exists an $M_6 > 0$ such that $\left\|\boldsymbol{\alpha}^l - \tilde{\boldsymbol{\alpha}}^l\right\|_2 \leq \frac{M_6}{\sqrt[4]{\tilde{d}}}\left\|\theta - \tilde{\theta}\right\|_2$ for all $l \geq 1$ (note that this is also true for $l = L+1$).

Thus, if (86) and (87) are true, then for all $\theta, \tilde{\theta} \in B(\theta^{(0)}, C_0)$, any $\boldsymbol{x}$ such that $\|\boldsymbol{x}\|_2 \leq 1$, we have

$$\begin{aligned}\left\|\nabla_{W^0} f(\boldsymbol{x}) - \nabla_{\tilde{W}^0} \tilde{f}(\boldsymbol{x})\right\|_2 &= \frac{1}{\sqrt{d_0}} \left\|\boldsymbol{x}\boldsymbol{\alpha}^{1\top} - \boldsymbol{x}\tilde{\boldsymbol{\alpha}}^{1\top}\right\|_2 \\ &\leq \frac{1}{\sqrt{d_0}} \left\|\boldsymbol{\alpha}^1 - \tilde{\boldsymbol{\alpha}}^1\right\|_2 \\ &\leq \frac{1}{\sqrt{d_0}} \frac{M_6}{\sqrt[4]{\tilde{d}}} \left\|\theta - \tilde{\theta}\right\|_2\end{aligned} \tag{91}$$

and for $l = 1, \cdots, L$, we have

$$\begin{aligned}\left\|\nabla_{W^l} f(\boldsymbol{x}) - \nabla_{\tilde{W}^l} \tilde{f}(\boldsymbol{x})\right\|_2 &= \frac{1}{\sqrt{\tilde{d}}} \left\|\boldsymbol{x}^l \boldsymbol{\alpha}^{l+1\top} - \tilde{\boldsymbol{x}}^l \tilde{\boldsymbol{\alpha}}^{l+1\top}\right\|_2 \\ &\leq \frac{1}{\sqrt{\tilde{d}}} \left(\left\|\boldsymbol{x}^l\right\|_2 \left\|\boldsymbol{\alpha}^{l+1} - \tilde{\boldsymbol{\alpha}}^{l+1}\right\|_2 + \left\|\boldsymbol{x}^l - \tilde{\boldsymbol{x}}^l\right\|_2 \left\|\tilde{\boldsymbol{\alpha}}^{l+1}\right\|_2\right) \\ &\leq \left(\frac{M_2 M_6}{\sqrt[4]{\tilde{d}}} + \frac{M_5 M_3}{\sqrt{\tilde{d}}}\right) \left\|\theta - \tilde{\theta}\right\|_2\end{aligned} \tag{92}$$

Moreover, for any $l = 0, \cdots, L$, there is

$$\left\|\nabla_{b^l} f(\boldsymbol{x}) - \nabla_{\tilde{b}^l} \tilde{f}(\boldsymbol{x})\right\|_2 = \beta \left\|\boldsymbol{\alpha}^{l+1} - \tilde{\boldsymbol{\alpha}}^{l+1}\right\|_2 \leq \frac{\beta M_6}{\sqrt[4]{\tilde{d}}} \left\|\theta - \tilde{\theta}\right\|_2 \tag{93}$$

Overall, we can see that there exists a constant $M_7 > 0$ such that $\left\|\nabla_\theta f(\boldsymbol{x}) - \nabla_{\tilde{\theta}} \tilde{f}(\boldsymbol{x})\right\|_2 \leq \frac{M_7}{\sqrt{n} \cdot \sqrt[4]{\tilde{d}}} \left\|\theta - \tilde{\theta}\right\|_2$, so that $\left\|J(\theta) - J(\tilde{\theta})\right\|_F \leq \frac{M_7}{\sqrt[4]{\tilde{d}}} \left\|\theta - \tilde{\theta}\right\|_2$. □

D.3.4 Proof of Theorem 4

First of all, for a linearized neural network (11), if we view $\{\nabla_\theta f^{(0)}(\boldsymbol{x}_i)\}_{i=1}^n$ as the inputs and $\{y_i - f^{(0)}(\boldsymbol{x}_i) + \langle \theta^{(0)}, \nabla_\theta f^{(0)}(\boldsymbol{x}_i)\rangle\}_{i=1}^n$ as the targets, then the model becomes a linear model. So by Theorem 2 we have the following corollary:

Corollary 15. *If $\nabla_\theta f^{(0)}(\boldsymbol{x}_1), \cdots, \nabla_\theta f^{(0)}(\boldsymbol{x}_n)$ are linearly independent, then there exists $\eta_0 > 0$ such that for any GRW satisfying Assumption 1, and any $\eta \leq \eta_0$, $\theta^{(t)}$ converges to the same interpolator θ^* that does not depend on q_i.*

Let $\eta_1 = \min\{\eta_0, \eta^*\}$, where η_0 is defined in Corollary 15 and η^* is defined in Lemma 5. Let $f_{\text{lin}}^{(t)}(\boldsymbol{x})$ and $f_{\text{linERM}}^{(t)}(\boldsymbol{x})$ be the linearized neural networks of $f^{(t)}(\boldsymbol{x})$ and $f_{\text{ERM}}^{(t)}(\boldsymbol{x})$, respectively. By Lemma 5, for any $\delta > 0$, there exists $\tilde{D} > 0$ and a constant C such that

$$\begin{cases} \sup\limits_{t \geq 0} \left|f_{\text{lin}}^{(t)}(\boldsymbol{x}) - f^{(t)}(\boldsymbol{x})\right| \leq C\tilde{d}^{-1/4} \\ \sup\limits_{t \geq 0} \left|f_{\text{linERM}}^{(t)}(\boldsymbol{x}) - f_{\text{ERM}}^{(t)}(\boldsymbol{x})\right| \leq C\tilde{d}^{-1/4} \end{cases} \tag{94}$$

By Corollary 15, we have

$$\lim_{t \to \infty} \left|f_{\text{lin}}^{(t)}(\boldsymbol{x}) - f_{\text{linERM}}^{(t)}(\boldsymbol{x})\right| = 0 \tag{95}$$

Summing the above yields

$$\limsup_{t \to \infty} \left|f^{(t)}(\boldsymbol{x}) - f_{\text{ERM}}^{(t)}(\boldsymbol{x})\right| \leq 2C\tilde{d}^{-1/4} \tag{96}$$

which is the result we want. □

D.4 Proofs for Subsection 4.3

D.4.1 A New Approximation Theorem

Lemma 16 (Approximation Theorem for Regularized GRW)**.** *For a wide fully-connected neural network f, denote $J(\theta) = \nabla_\theta f(\boldsymbol{X}; \theta) \in \mathbb{R}^{p\times n}$ and $g(\theta) = \nabla_{\hat{y}} \ell(f(\boldsymbol{X}; \theta), \boldsymbol{Y}) \in \mathbb{R}^n$. Given that the loss function ℓ satisfies: $\nabla_\theta g(\theta) = J(\theta)U(\theta)$ for any θ, and $U(\theta)$ is a positive semi-definite diagonal matrix whose elements are uniformly bounded, we have: for any GRW that minimizes the regularized weighted empirical risk (13) with a sufficiently small learning rate η, there is: for a sufficiently large $\tilde{d}$, with high probability over random initialization, on any test point $\boldsymbol{x}$ such that $\|\boldsymbol{x}\|_2 \le 1$,*

$$\sup_{t\ge 0} \left| f_{\text{linreg}}^{(t)}(\boldsymbol{x}) - f_{\text{reg}}^{(t)}(\boldsymbol{x}) \right| \le C\tilde{d}^{-1/4} \tag{97}$$

where both $f_{\text{linreg}}^{(t)}$ and $f_{\text{reg}}^{(t)}$ are trained by the same regularized GRW and start from the same initial point.

First of all, with some simple linear algebra analysis, we can prove the following proposition:

Proposition 17. *For any positive definite symmetric matrix $\boldsymbol{H} \in \mathbb{R}^{n\times n}$, denote its largest and smallest eigenvalues by $\lambda^{\max}$ and $\lambda^{\min}$. Then, for any positive semi-definite diagonal matrix $\boldsymbol{Q} = \text{diag}(q_1, \cdots, q_n)$, $\boldsymbol{HQ}$ has n eigenvalues that all lie in $[\min_i q_i \cdot \lambda^{\min}, \max_i q_i \cdot \lambda^{\max}]$.*

Proof. $\boldsymbol{H}$ is a positive definite symmetric matrix, so there exists $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ such that $\boldsymbol{H} = \boldsymbol{A}^\top \boldsymbol{A}$, and $\boldsymbol{A}$ is full-rank. First, any eigenvalue of $\boldsymbol{AQA}^\top$ is also an eigenvalue of $\boldsymbol{A}^\top \boldsymbol{AQ}$, because for any eigenvalue λ of $\boldsymbol{AQA}^\top$ we have some $\boldsymbol{v} \neq 0$ such that $\boldsymbol{AQA}^\top \boldsymbol{v} = \lambda \boldsymbol{v}$. Multiplying both sides by $\boldsymbol{A}^\top$ on the left yields $\boldsymbol{A}^\top \boldsymbol{AQ}(\boldsymbol{A}^\top \boldsymbol{v}) = \lambda(\boldsymbol{A}^\top \boldsymbol{v})$ which implies that λ is also an eigenvalue of $\boldsymbol{A}^\top \boldsymbol{AQ}$ because $\boldsymbol{A}^\top \boldsymbol{v} \neq 0$ as $\lambda \boldsymbol{v} \neq 0$.

Second, by condition we know that the eigenvalues of $\boldsymbol{A}^\top \boldsymbol{A}$ are all in $[\lambda^{\min}, \lambda^{\max}]$ where $\lambda^{\min} > 0$, which implies for any unit vector $\boldsymbol{v}$, $\boldsymbol{v}^\top \boldsymbol{A}^\top \boldsymbol{Av} \in [\lambda^{\min}, \lambda^{\max}]$, which is equivalent to $\|\boldsymbol{Av}\|_2 \in [\sqrt{\lambda^{\min}}, \sqrt{\lambda^{\max}}]$. Thus, we have $\boldsymbol{v}^\top \boldsymbol{A}^\top \boldsymbol{QAv} \in [\lambda^{\min} \min_i q_i, \lambda^{\max} \max_i q_i]$, which implies that the eigenvalues of $\boldsymbol{A}^\top \boldsymbol{QA}$ are all in $[\lambda^{\min} \min_i q_i, \lambda^{\max} \max_i q_i]$.

Thus, the eigenvalues of $\boldsymbol{HQ} = \boldsymbol{A}^\top \boldsymbol{AQ}$ are all in $[\lambda^{\min} \min_i q_i, \lambda^{\max} \max_i q_i]$. □

Proof of Lemma 16 By the condition ℓ satisfies, without loss of generality, assume that the elements of $U(\theta)$ are in $[0, 1]$ for all θ. Then, let $\eta \le (\mu + \lambda^{\min} + \lambda^{\max})^{-1}$. (If the elements of $U(\theta)$ are bounded by $[0, C]$, then we can let $\eta \le (\mu + C\lambda^{\min} + C\lambda^{\max})^{-1}$ and prove the result in the same way.)

With L_2 penalty, the update rule of the GRW for the neural network is:

$$\theta^{(t+1)} = \theta^{(t)} - \eta J(\theta^{(t)}) \boldsymbol{Q}^{(t)} g(\theta^{(t)}) - \eta\mu(\theta^{(t)} - \theta^{(0)}) \tag{98}$$

And the update rule for the linearized neural network is:

$$\theta_{\text{lin}}^{(t+1)} = \theta_{\text{lin}}^{(t)} - \eta J(\theta^{(0)}) \boldsymbol{Q}^{(t)} g(\theta_{\text{lin}}^{(t)}) - \eta\mu(\theta_{\text{lin}}^{(t)} - \theta^{(0)}) \tag{99}$$

By Proposition 11, $f(\boldsymbol{x}; \theta)$ converges in probability to a zero-mean Gaussian process. Thus, for any $\delta > 0$, there exists a constant $R_0 > 0$ such that with probability at least $(1 - \delta/3)$, $\left\|g(\theta^{(0)})\right\|_2 < R_0$. Let M be as defined in Lemma 13. Denote $A = \eta M R_0$, and let $C_0 = \frac{4A}{\eta\mu}$ in Lemma 13[1]. By Lemma 13, there exists D_1 such that for all $\tilde{d} \ge D_1$, with probability at least $(1 - \delta/3)$, (52) is true.

Similar to the proof of Proposition 17, we can show that for arbitrary $\tilde{\theta}$, all non-zero eigenvalues of $J(\theta^{(0)}) \boldsymbol{Q}^{(t)} U(\tilde{\theta}) J(\theta^{(0)})^\top$ are eigenvalues of $J(\theta^{(0)})^\top J(\theta^{(0)}) \boldsymbol{Q}^{(t)} U(\tilde{\theta})$. This is because for any $\lambda \neq 0$, if $J(\theta^{(0)}) \boldsymbol{Q}^{(t)} U(\tilde{\theta}) J(\theta^{(0)})^\top \boldsymbol{v} = \lambda \boldsymbol{v}$, then $J(\theta^{(0)})^\top J(\theta^{(0)}) \boldsymbol{Q}^{(t)} U(\tilde{\theta})(J(\theta^{(0)})^\top \boldsymbol{v}) = \lambda(J(\theta^{(0)})^\top \boldsymbol{v})$, and $J(\theta^{(0)})^\top \boldsymbol{v} \neq 0$ since $\lambda \boldsymbol{v} \neq 0$, so λ is also an eigenvalue of

[1] Note that Lemma 13 only depends on the network structure and does not depend on the update rule, so we can use this lemma here.

$J(\theta^{(0)})^\top J(\theta^{(0)})\boldsymbol{Q}^{(t)}U(\tilde{\theta})$. On the other hand, by Proposition 3, $J(\theta^{(0)})^\top J(\theta^{(0)})\boldsymbol{Q}^{(t)}U(\tilde{\theta})$ converges in probability to $\Theta\boldsymbol{Q}^{(t)}U(\tilde{\theta})$ whose eigenvalues are all in $[0, \lambda^{\max}]$ by Proposition 17. So there exists D_2 such that for all $\tilde{d} \geq D_2$, with probability at least $(1-\delta/3)$, the eigenvalues of $J(\theta^{(0)})\boldsymbol{Q}^{(t)}U(\tilde{\theta})J(\theta^{(0)})^\top$ are all in $[0, \lambda^{\max}+\lambda^{\min}]$ for all t.

By union bound, with probability at least $(1-\delta)$, all three above are true, which we will assume in the rest of this proof.

First, we need to prove that there exists D_0 such that for all $\tilde{d} \geq D_0$, $\sup_{t\geq 0}\left\|\theta^{(t)}-\theta^{(0)}\right\|_2$ is bounded with high probability. Denote $a_t = \theta^{(t)}-\theta^{(0)}$. By (98) we have

$$\begin{aligned} a_{t+1} =&(1-\eta\mu)a_t - \eta[J(\theta^{(t)})-J(\theta^{(0)})]\boldsymbol{Q}^{(t)}g(\theta^{(t)}) \\ &- \eta J(\theta^{(0)})\boldsymbol{Q}^{(t)}[g(\theta^{(t)})-g(\theta^{(0)})] - \eta J(\theta^{(0)})\boldsymbol{Q}^{(t)}g(\theta^{(0)}) \end{aligned} \tag{100}$$

which implies

$$\begin{aligned} \|a_{t+1}\|_2 \leq &\left\|(1-\eta\mu)\boldsymbol{I} - \eta J(\theta^{(0)})\boldsymbol{Q}^{(t)}U(\tilde{\theta}^{(t)})J(\tilde{\theta}^{(t)})^\top\right\|_2 \|a_t\|_2 \\ &+ \eta\left\|J(\theta^{(t)})-J(\theta^{(0)})\right\|_F\left\|g(\theta^{(t)})\right\|_2 + \eta\left\|J(\theta^{(0)})\right\|_F\left\|g(\theta^{(0)})\right\|_2 \end{aligned} \tag{101}$$

where $\tilde{\theta}^{(t)}$ is some linear interpolation between $\theta^{(t)}$ and $\theta^{(0)}$. Our choice of η ensures that $\eta\mu < 1$.

Now we prove by induction that $\|a_t\|_2 < C_0$. It is true for $t=0$, so we need to prove that if $\|a_t\|_2 < C_0$, then $\|a_{t+1}\|_2 < C_0$.

For the first term on the right-hand side of (101), we have

$$\begin{aligned} \left\|(1-\eta\mu)\boldsymbol{I} - \eta J(\theta^{(0)})\boldsymbol{Q}^{(t)}U(\tilde{\theta}^{(t)})J(\tilde{\theta}^{(t)})^\top\right\|_2 \leq &(1-\eta\mu)\left\|\boldsymbol{I} - \frac{\eta}{1-\eta\mu}J(\theta^{(0)})\boldsymbol{Q}^{(t)}U(\tilde{\theta}^{(t)})J(\theta^{(0)})^\top\right\|_2 \\ &+ \eta\left\|J(\theta^{(0)})\right\|_F\left\|J(\tilde{\theta}^{(t)})-J(\theta^{(0)})\right\|_F \end{aligned} \tag{102}$$

Since $\eta/(1-\eta\mu) \leq (\lambda^{\min}+\lambda^{\max})^{-1}$ by our choice of η, we have

$$\left\|\boldsymbol{I} - \frac{\eta}{1-\eta\mu}J(\theta^{(0)})\boldsymbol{Q}^{(t)}U(\tilde{\theta}^{(t)})J(\theta^{(0)})^\top\right\|_2 \leq 1 \tag{103}$$

On the other hand, we can use (52) since $\|a_t\|_2 < C_0$, so $\left\|J(\theta^{(0)})\right\|_F\left\|J(\tilde{\theta}^{(t)})-J(\theta^{(0)})\right\|_F \leq \frac{M^2}{\sqrt[4]{\tilde{d}}}C_0$. Therefore, there exists D_3 such that for all $\tilde{d} \geq D_3$,

$$\left\|(1-\eta\mu)\boldsymbol{I} - \eta J(\theta^{(0)})\boldsymbol{Q}^{(t)}U(\tilde{\theta}^{(t)})J(\tilde{\theta}^{(t)})^\top\right\|_2 \leq 1 - \frac{\eta\mu}{2} \tag{104}$$

For the second term, we have

$$\begin{aligned} \left\|g(\theta^{(t)})\right\|_2 &\leq \left\|g(\theta^{(t)})-g(\theta^{(0)})\right\|_2 + \left\|g(\theta^{(0)})\right\|_2 \\ &\leq \left\|J(\tilde{\theta}^{(t)})\right\|_2\left\|U(\tilde{\theta}^{(t)})\right\|_2\left\|\theta^{(t)}-\theta^{(0)}\right\|_2 + R_0 \leq MC_0 + R_0 \end{aligned} \tag{105}$$

And for the third term, we have

$$\eta\left\|J(\theta^{(0)})\right\|_F\left\|g(\theta^{(0)})\right\|_2 \leq \eta M R_0 = A \tag{106}$$

Thus, we have

$$\|a_{t+1}\|_2 \leq \left(1-\frac{\eta\mu}{2}\right)\|a_t\|_2 + \frac{\eta M(MC_0+R_0)}{\sqrt[4]{\tilde{d}}} + A \tag{107}$$

So there exists D_4 such that for all $\tilde{d} \geq D_4$, $\|a_{t+1}\|_2 \leq \left(1-\frac{\eta\mu}{2}\right)\|a_t\|_2 + 2A$. This shows that if $\|a_t\|_2 < C_0$ is true, then $\|a_{t+1}\|_2 < C_0$ will also be true.

In conclusion, for all $\tilde{d} \geq D_0 = \max\{D_1, D_2, D_3, D_4\}$, $\left\|\theta^{(t)} - \theta^{(0)}\right\|_2 < C_0$ is true for all t. This also implies that for $C_1 = MC_0 + R_0$, we have $\left\|g(\theta^{(t)})\right\|_2 \leq C_1$ for all t by (105). Similarly, we can prove that $\|\theta_{\text{lin}}^{(t)} - \theta^{(0)}\|_2 < C_0$ for all t.

Second, let $\Delta_t = \theta_{\text{lin}}^{(t)} - \theta^{(t)}$. Then we have

$$\Delta_{t+1} - \Delta_t = \eta(J(\theta^{(t)})\boldsymbol{Q}^{(t)}g(\theta^{(t)}) - J(\theta^{(0)})\boldsymbol{Q}^{(t)}g(\theta_{\text{lin}}^{(t)}) - \mu\Delta_t) \tag{108}$$

which implies

$$\Delta_{t+1} = \left[(1-\eta\mu)\boldsymbol{I} - \eta J(\theta^{(0)})\boldsymbol{Q}^{(t)}U(\tilde{\theta}^{(t)})J(\tilde{\theta}^{(t)})^\top\right]\Delta_t + \eta(J(\theta^{(t)}) - J(\theta^{(0)}))\boldsymbol{Q}^{(t)}g(\theta^{(t)}) \tag{109}$$

where $\tilde{\theta}^{(t)}$ is some linear interpolation between $\theta^{(t)}$ and $\theta_{\text{lin}}^{(t)}$. By (104), with probability at least $(1-\delta)$ for all $\tilde{d} \geq D_0$, we have

$$\begin{aligned}\|\Delta_{t+1}\|_2 &\leq \left\|(1-\eta\mu)\boldsymbol{I} - \eta J(\theta^{(0)})\boldsymbol{Q}^{(t)}U(\tilde{\theta}^{(t)})J(\tilde{\theta}^{(t)})^\top\right\|_2 \|\Delta_t\|_2 + \eta\left\|J(\theta^{(t)}) - J(\theta^{(0)})\right\|_F \left\|g(\theta^{(t)})\right\|_2 \\ &\leq \left(1 - \frac{\eta\mu}{2}\right)\|\Delta_t\|_2 + \eta\frac{M}{\sqrt[4]{\tilde{d}}}C_0C_1\end{aligned} \tag{110}$$

Again, as $\Delta_0 = 0$, we can prove by induction that for all t,

$$\|\Delta_t\|_2 < \frac{2MC_0C_1}{\mu}\tilde{d}^{-1/4} \tag{111}$$

For any test point $\boldsymbol{x}$ such that $\|\boldsymbol{x}\|_2 \leq 1$, we have

$$\begin{aligned}\left|f_{\text{reg}}^{(t)}(\boldsymbol{x}) - f_{\text{linreg}}^{(t)}(\boldsymbol{x})\right| &= \left|f(\boldsymbol{x};\theta^{(t)}) - f_{\text{lin}}(\boldsymbol{x};\theta_{\text{lin}}^{(t)})\right| \\ &\leq \left|f(\boldsymbol{x};\theta^{(t)}) - f_{\text{lin}}(\boldsymbol{x};\theta^{(t)})\right| + \left|f_{\text{lin}}(\boldsymbol{x};\theta^{(t)}) - f_{\text{lin}}(\boldsymbol{x};\theta_{\text{lin}}^{(t)})\right| \\ &\leq \left|f(\boldsymbol{x};\theta^{(t)}) - f_{\text{lin}}(\boldsymbol{x};\theta^{(t)})\right| + \left\|\nabla_\theta f(\boldsymbol{x};\theta^{(0)})\right\|_2\left\|\theta^{(t)} - \theta_{\text{lin}}^{(t)}\right\|_2 \\ &\leq \left|f(\boldsymbol{x};\theta^{(t)}) - f_{\text{lin}}(\boldsymbol{x};\theta^{(t)})\right| + M\|\Delta_t\|_2\end{aligned} \tag{112}$$

For the first term, note that

$$\begin{cases} f(\boldsymbol{x};\theta^{(t)}) - f(\boldsymbol{x};\theta^{(0)}) = \nabla_\theta f(\boldsymbol{x};\tilde{\theta}^{(t)})(\theta^{(t)} - \theta^{(0)}) \\ f_{\text{lin}}(\boldsymbol{x};\theta^{(t)}) - f_{\text{lin}}(\boldsymbol{x};\theta^{(0)}) = \nabla_\theta f(\boldsymbol{x};\theta^{(0)})(\theta^{(t)} - \theta^{(0)}) \end{cases} \tag{113}$$

where $\tilde{\theta}^{(t)}$ is some linear interpolation between $\theta^{(t)}$ and $\theta^{(0)}$. Since $f(\boldsymbol{x};\theta^{(0)}) = f_{\text{lin}}(\boldsymbol{x};\theta^{(0)})$,

$$\left|f(\boldsymbol{x};\theta^{(t)}) - f_{\text{lin}}(\boldsymbol{x};\theta^{(t)})\right| \leq \left\|\nabla_\theta f(\boldsymbol{x};\tilde{\theta}^{(t)}) - \nabla_\theta f(\boldsymbol{x};\theta^{(0)})\right\|_2\left\|\theta^{(t)} - \theta^{(0)}\right\|_2 \leq \frac{M}{\sqrt[4]{\tilde{d}}}C_0^2 \tag{114}$$

Thus, we have shown that for all $\tilde{d} \geq D_0$, with probability at least $(1-\delta)$ for all t and all $\boldsymbol{x}$,

$$\left|f_{\text{reg}}^{(t)}(\boldsymbol{x}) - f_{\text{linreg}}^{(t)}(\boldsymbol{x})\right| \leq \left(MC_0^2 + \frac{2M^2C_0C_1}{\mu}\right)\tilde{d}^{-1/4} = O(\tilde{d}^{-1/4}) \tag{115}$$

which is the result we need. □

D.4.2 Result for Linearized Neural Networks

Lemma 18. *Suppose there exists $M_0 > 0$ such that $\left\|\nabla_\theta f^{(0)}(\boldsymbol{x})\right\|_2 \leq M_0$ for all test point $\boldsymbol{x}$. If the gradients $\nabla_\theta f^{(0)}(\boldsymbol{x}_1), \cdots, \nabla_\theta f^{(0)}(\boldsymbol{x}_n)$ are linearly independent, and the empirical training risk of $f_{\text{linreg}}^{(t)}$ satisfies*

$$\limsup_{t\to\infty}\hat{\mathcal{R}}(f_{\text{linreg}}^{(t)}) < \epsilon, \tag{116}$$

for some $\epsilon > 0$, then for $\boldsymbol{x}$ such that $\|\boldsymbol{x}\|_2 \leq 1$ we have

$$\limsup_{t\to\infty}\left|f_{\text{linreg}}^{(t)}(\boldsymbol{x}) - f_{\text{linERM}}^{(t)}(\boldsymbol{x})\right| = O(\sqrt{\epsilon}). \tag{117}$$

First, we can see that under the new weight update rule, $\theta^{(t)} - \theta^{(0)} \in \text{span}\{\nabla_\theta f^{(0)}(\boldsymbol{x}_1), \cdots, \nabla_\theta f^{(0)}(\boldsymbol{x}_n)\}$ is still true for all t. Let θ^* be the interpolator in $\text{span}(\nabla_\theta f^{(0)}(\boldsymbol{x}_1), \cdots, \nabla_\theta f^{(0)}(\boldsymbol{x}_n))$, then the empirical risk of θ is $\frac{1}{2n}\sum_{i=1}^n \langle \theta - \theta^*, \nabla_\theta f^{(0)}(\boldsymbol{x}_i)\rangle^2 = \frac{1}{2n}\left\|\nabla_\theta f^{(0)}(\boldsymbol{X})^\top(\theta - \theta^*)\right\|_2^2$. Thus, there exists $T > 0$ such that for any $t \geq T$,

$$\left\|\nabla_\theta f^{(0)}(\boldsymbol{X})^\top(\theta^{(t)} - \theta^*)\right\|_2^2 \leq 2n\epsilon \tag{118}$$

Let the smallest singular value of $\frac{1}{\sqrt{n}}\nabla_\theta f^{(0)}(\boldsymbol{X})$ be $s^{\min}$, and we have $s^{\min} > 0$. Note that the column space of $\nabla_\theta f^{(0)}(\boldsymbol{X})$ is exactly $\text{span}(\nabla_\theta f^{(0)}(\boldsymbol{x}_1), \cdots, \nabla_\theta f^{(0)}(\boldsymbol{x}_n))$. Define $\boldsymbol{H} \in \mathbb{R}^{p\times n}$ such that its columns form an orthonormal basis of this subspace, then there exists $\boldsymbol{G} \in R^{n\times n}$ such that $\nabla_\theta f^{(0)}(\boldsymbol{X}) = \boldsymbol{H}\boldsymbol{G}$, and the smallest singular value of $\frac{1}{\sqrt{n}}\boldsymbol{G}$ is also $s^{\min}$. Since $\theta^{(t)} - \theta^{(0)}$ is also in this subspace, there exists $\boldsymbol{v} \in \mathbb{R}^n$ such that $\theta^{(t)} - \theta^* = \boldsymbol{H}\boldsymbol{v}$. Then we have $\sqrt{2n\epsilon} \geq \left\|\boldsymbol{G}^\top\boldsymbol{H}^\top\boldsymbol{H}\boldsymbol{v}\right\|_2 = \left\|\boldsymbol{G}^\top\boldsymbol{v}\right\|_2$. Thus, $\|\boldsymbol{v}\|_2 \leq \frac{\sqrt{2\epsilon}}{s^{\min}}$, which implies

$$\left\|\theta^{(t)} - \theta^*\right\|_2 \leq \frac{\sqrt{2\epsilon}}{s^{\min}} \tag{119}$$

We have already proved in previous results that if we minimize the unregularized risk with ERM, then θ always converges to the interpolator θ^*. So for any $t \geq T$ and any test point $\boldsymbol{x}$ such that $\|\boldsymbol{x}\|_2 \leq 1$, we have

$$|f_{\text{linreg}}^{(t)}(\boldsymbol{x}) - f_{\text{linERM}}^{(t)}(\boldsymbol{x})| = |\langle \theta^{(t)} - \theta^*, \nabla_\theta f^{(0)}(\boldsymbol{x})\rangle| \leq \frac{M_0\sqrt{2\epsilon}}{s^{\min}} \tag{120}$$

which implies (117). □

D.4.3 Proof of Theorem 6

Given that $\hat{\mathcal{R}}(f_{\text{linreg}}^{(t)}) < \epsilon$ for sufficiently large t, Lemma 16 implies that

$$\left|\hat{\mathcal{R}}(f_{\text{linreg}}^{(t)}) - \hat{\mathcal{R}}(f_{\text{reg}}^{(t)})\right| = O(\tilde{d}^{-1/4}\sqrt{\epsilon} + \tilde{d}^{-1/2}) \tag{121}$$

So for a fixed ϵ, there exists $D > 0$ such that for all $d \geq D$, for sufficiently large t,

$$\hat{\mathcal{R}}(f_{\text{reg}}^{(t)}) < \epsilon \Rightarrow \hat{\mathcal{R}}(f_{\text{linreg}}^{(t)}) < 2\epsilon \tag{122}$$

By Lemma 5 and Lemma 16, we have

$$\begin{cases} \sup\limits_{t\geq 0}\left|f_{\text{linERM}}^{(t)}(\boldsymbol{x}) - f_{\text{ERM}}^{(t)}(\boldsymbol{x})\right| = O(\tilde{d}^{-1/4}) \\ \sup\limits_{t\geq 0}\left|f_{\text{linreg}}^{(t)}(\boldsymbol{x}) - f_{\text{reg}}^{(t)}(\boldsymbol{x})\right| = O(\tilde{d}^{-1/4}) \end{cases} \tag{123}$$

Combining Lemma 18 with (123) derives

$$\limsup_{t\to\infty}\left|f_{\text{reg}}^{(t)}(\boldsymbol{x}) - f_{\text{ERM}}^{(t)}(\boldsymbol{x})\right| = O(\tilde{d}^{-1/4} + \sqrt{\epsilon}) \tag{124}$$

Letting $\tilde{d} \to \infty$ leads to the result we need. □

Remark. One might wonder whether $\|\nabla_\theta f^{(0)}(\boldsymbol{x})\|_2$ will diverge as $\tilde{d} \to \infty$. In fact, in Lemma 13, we have proved that there exists a constant M such that with high probability, for any $\tilde{d}$ there is $\|\nabla_\theta f^{(0)}(\boldsymbol{x})\|_2 \leq M$ for any $\boldsymbol{x}$ such that $\|\boldsymbol{x}\|_2 \leq 1$. Therefore, it is fine to suppose that there exists such an M_0.

D.5 Proofs for Subsection 5.1

D.5.1 Proof of Theorem 7

First we need to show that $\hat{\theta}_{\text{MM}}$ is unique. Suppose both θ_1 and θ_2 maximize $\min_{i=1,\cdots,n} y_i \cdot \langle \theta, \boldsymbol{x}_i \rangle$ and $\theta_1 \neq \theta_2$, $\|\theta_1\|_2 = \|\theta_2\|_2 = 1$. Then consider $\theta_0 = \theta/\|\theta\|_2$ where $\theta = (\theta_1 + \theta_2)/2$. Obviously, $\|\theta\|_2 < 1$, and for any i, $y_i \cdot \langle \theta, \boldsymbol{x}_i \rangle = (y_i \cdot \langle \theta_1, \boldsymbol{x}_i \rangle + y_i \cdot \langle \theta_2, \boldsymbol{x}_i \rangle)/2$, so $y_i \cdot \langle \theta_0, \boldsymbol{x}_i \rangle > \min\{y_i \cdot \langle \theta_1, \boldsymbol{x}_i \rangle, y_i \cdot \langle \theta_2, \boldsymbol{x}_i \rangle\}$, which implies that $\min_{i=1,\cdots,n} y_i \cdot \langle \theta_0, \boldsymbol{x}_i \rangle > \min\{\min_{i=1,\cdots,n} y_i \cdot \langle \theta_1, \boldsymbol{x}_i \rangle, \min_{i=1,\cdots,n} y_i \cdot \langle \theta_2, \boldsymbol{x}_i \rangle\}$, contradiction!

Now we start proving the result. Without loss of generality, let $(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_m, y_m)$ be the samples with the smallest margin to $\boldsymbol{u}$, i.e.

$$\arg\min_{1 \leq i \leq n} y_i \cdot \langle \boldsymbol{u}, \boldsymbol{x}_i \rangle = \{1, \cdots, m\} \tag{125}$$

And denote $y_1 \cdot \langle \boldsymbol{u}, \boldsymbol{x}_1 \rangle = \cdots = y_m \cdot \langle \boldsymbol{u}, \boldsymbol{x}_m \rangle = \gamma_{\boldsymbol{u}}$. Since the training error converges to 0, $\gamma_{\boldsymbol{u}} > 0$. Note that for the logistic loss, if $y_i \cdot \langle \theta, \boldsymbol{x}_i \rangle < y_j \cdot \langle \theta, \boldsymbol{x}_j \rangle$, then for any $M > 0$, there exists an $R_M > 0$ such that for all $R \geq R_M$,

$$\frac{\nabla_\theta \ell(\langle R\theta, \boldsymbol{x}_i \rangle, y_i)}{\nabla_\theta \ell(\langle R\theta, \boldsymbol{x}_j \rangle, y_j)} > M \tag{126}$$

which can be shown with some simple calculation. And because the training error converges to 0, we must have $\left\|\theta^{(t)}\right\| \to \infty$. Then, by Assumption 3 this means that when t gets sufficiently large, the impact of $(\boldsymbol{x}_j, y_j)$ to $\theta^{(t)}$ where $j > m$ is an infinitesimal compared to $(\boldsymbol{x}_i, y_i)$ where $i \leq m$ (because there exists a positive constant δ such that $q_i^{(t)} > \delta$ for all sufficiently large t by Assumption 3). Thus, we must have $\boldsymbol{u} \in \text{span}\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_m\}$.

Let $\boldsymbol{u} = \alpha_1 y_1 \boldsymbol{x}_1 + \cdots + \alpha_m y_m \boldsymbol{x}_m$. Now we show that $\alpha_i \geq 0$ for all $i = 1, \cdots, m$. This is because when t is sufficiently large such that the impact of $(\boldsymbol{x}_j, y_j)$ to $\theta^{(t)}$ where $j > m$ becomes infinitesimal, we have

$$\theta^{(t+1)} - \theta^{(t)} \approx \eta \frac{q_i^{(t)} \exp(y_i \cdot \langle \theta^{(t)}, \boldsymbol{x}_i \rangle)}{1 + \exp(y_i \cdot \langle \theta^{(t)}, \boldsymbol{x}_i \rangle)} y_i \boldsymbol{x}_i \tag{127}$$

and since $\|\theta^{(t)}\| \to \infty$ as $t \to \infty$, we have

$$\alpha_i \propto \lim_{T \to \infty} \sum_{t=T_0}^{T} \frac{q_i^{(t)} \exp(y_i \cdot \langle \theta^{(t)}, \boldsymbol{x}_i \rangle)}{1 + \exp(y_i \cdot \langle \theta^{(t)}, \boldsymbol{x}_i \rangle)} := \lim_{T \to \infty} \alpha_i(T) \tag{128}$$

where T_0 is sufficiently large. Here the notion $\alpha_i \propto \lim_{T \to \infty} \alpha_i(T)$ means that $\lim_{T \to \infty} \frac{\alpha_i(T)}{\alpha_j(T)} = \frac{\alpha_i}{\alpha_j}$ for any pair of i, j and $\alpha_j \neq 0$. Note that each term in the sum is non-negative. This implies that all $\alpha_1, \cdots, \alpha_m$ have the same sign (or equal to 0). On the other hand,

$$\sum_{i=1}^{m} \alpha_i \gamma_{\boldsymbol{u}} = \sum_{i=1}^{m} \alpha_i y_i \cdot \langle \boldsymbol{u}, \boldsymbol{x}_i \rangle = \langle \boldsymbol{u}, \boldsymbol{u} \rangle > 0 \tag{129}$$

Thus, $\alpha_i \geq 0$ for all i and at least one of them is positive. Now suppose $\boldsymbol{u} \neq \hat{\theta}_{\text{MM}}$, which means that $\gamma_{\boldsymbol{u}}$ is smaller than the margin of $\hat{\theta}_{\text{MM}}$. Then, for all $i = 1, \cdots, m$, there is $y_i \cdot \langle \boldsymbol{u}, \boldsymbol{x}_i \rangle < y_i \cdot \langle \hat{\theta}_{\text{MM}}, \boldsymbol{x}_i \rangle$. This implies that

$$\langle \boldsymbol{u}, \boldsymbol{u} \rangle = \sum_{i=1}^{m} \alpha_i y_i \cdot \langle \boldsymbol{u}, \boldsymbol{x}_i \rangle < \sum_{i=1}^{m} \alpha_i y_i \cdot \langle \hat{\theta}_{\text{MM}}, \boldsymbol{x}_i \rangle = \langle \hat{\theta}_{\text{MM}}, \boldsymbol{u} \rangle \tag{130}$$

which is a contradiction. Thus, we must have $\boldsymbol{u} = \hat{\theta}_{\text{MM}}$. □

D.5.2 Proof of Theorem 8

Denote the largest and smallest eigenvalues of $\boldsymbol{X}^\top \boldsymbol{X}$ by $\lambda^{\max}$ and $\lambda^{\min}$, and by condition we have $\lambda^{\min} > 0$. Let $\epsilon = \min\{\frac{q^*}{3}, \frac{(q^* \lambda^{\min})^2}{192 \lambda^{\max 2}}\}$. Then similar to the proof in Appendix D.2.2, there exists t_ϵ

such that for all $t \geq t_\epsilon$ and all i, $q_i^{(t)} \in (q_i - \epsilon, q_i + \epsilon)$. Denote $\boldsymbol{Q} = \mathrm{diag}(q_1, \cdots, q_n)$, then for all $t \geq t_\epsilon$, $\boldsymbol{Q}^{(t)} := \boldsymbol{Q}_\epsilon^{(t)} = \sqrt{\boldsymbol{Q}}\sqrt{\boldsymbol{Q}_{3\epsilon}^{(t)}}$, where we use the subscript ϵ to indicate that $\left\|\boldsymbol{Q}_\epsilon^{(t)} - \boldsymbol{Q}\right\|_2 < \epsilon$.

First, we prove that $F(\theta)$ is L-smooth as long as $\|\boldsymbol{x}_i\|_2 \leq 1$ for all i. The gradient of F is

$$\nabla F(\theta) = \sum_{i=1}^n q_i \nabla_{\hat{y}} \ell(\langle \theta, \boldsymbol{x}_i \rangle, y_i) \boldsymbol{x}_i \tag{131}$$

Since $\ell(\hat{y}, y)$ is L-smooth in $\hat{y}$, we have for any θ_1, θ_2 and any i,

$$\begin{aligned}
&\ell(\langle \theta_2, \boldsymbol{x}_i \rangle, y_i) - \ell(\langle \theta_1, \boldsymbol{x}_i \rangle, y_i) \\
\leq& \nabla_{\hat{y}} \ell(\langle \theta_1, \boldsymbol{x}_i \rangle, y_i) \cdot (\langle \theta_2, \boldsymbol{x}_i \rangle - \langle \theta_1, \boldsymbol{x}_i \rangle) + \frac{L}{2} (\langle \theta_2, \boldsymbol{x}_i \rangle - \langle \theta_1, \boldsymbol{x}_i \rangle)^2 \\
=& \langle \nabla_{\hat{y}} \ell(\langle \theta_1, \boldsymbol{x}_i \rangle, y_i) \cdot \boldsymbol{x}_i, \theta_2 - \theta_1 \rangle + \frac{L}{2} (\langle \theta_2 - \theta_1, \boldsymbol{x}_i \rangle)^2 \\
\leq& \langle \nabla_{\hat{y}} \ell(\langle \theta_1, \boldsymbol{x}_i \rangle, y_i) \cdot \boldsymbol{x}_i, \theta_2 - \theta_1 \rangle + \frac{L}{2} \|\theta_2 - \theta_1\|_2^2
\end{aligned} \tag{132}$$

Thus, we have

$$\begin{aligned}
F(\theta_2) - F(\theta_1) =& \sum_{i=1}^n q_i \left[\ell(\langle \theta_2, \boldsymbol{x}_i \rangle, y_i) - \ell(\langle \theta_1, \boldsymbol{x}_i \rangle, y_i)\right] \\
\leq& \sum_{i=1}^n q_i \langle \nabla_{\hat{y}} \ell(\langle \theta_1, \boldsymbol{x}_i \rangle, y_i) \cdot \boldsymbol{x}_i, \theta_2 - \theta_1 \rangle + \frac{L}{2} \sum_{i=1}^n q_i \|\theta_2 - \theta_1\|_2^2 \\
=& \langle \nabla F(\theta_1), \theta_2 - \theta_1 \rangle + \frac{L}{2} \|\theta_2 - \theta_1\|_2^2
\end{aligned} \tag{133}$$

which implies that $F(\theta)$ is L-smooth.

Denote $\tilde{g}(\theta) = \nabla_{\hat{y}} \ell(f(\boldsymbol{X}; \theta), \boldsymbol{Y}) \in \mathbb{R}^n$, then $\nabla F(\theta^{(t)}) = \boldsymbol{X}\boldsymbol{Q}\tilde{g}(\theta^{(t)})$, and the update rule is

$$\theta^{(t+1)} = \theta^{(t)} - \eta \boldsymbol{X}\boldsymbol{Q}^{(t)} \tilde{g}(\theta^{(t)}) \tag{134}$$

So by the upper quadratic bound, we have

$$F(\theta^{(t+1)}) \leq F(\theta^{(t)}) - \eta \langle \boldsymbol{X}\boldsymbol{Q}\tilde{g}(\theta^{(t)}), \boldsymbol{X}\boldsymbol{Q}^{(t)}\tilde{g}(\theta^{(t)}) \rangle + \frac{\eta^2 L}{2} \left\|\boldsymbol{X}\boldsymbol{Q}^{(t)}\tilde{g}(\theta^{(t)})\right\|_2^2 \tag{135}$$

Let $\eta_1 = \frac{q^* \lambda^{\min}}{2L(1+3\epsilon)\lambda^{\max}}$. Similar to what we did in Appendix D.2.2 (Eqn. (40)), we can prove that for all $\eta \leq \eta_1$, (135) implies that for all $t \geq t_\epsilon$, there is

$$\begin{aligned}
F(\theta^{(t+1)}) &\leq F(\theta^{(t)}) - \frac{\eta q^* \lambda^{\min}}{2} \left\|\sqrt{\boldsymbol{Q}}\tilde{g}(\theta^{(t)})\right\|_2^2 + \frac{\eta^2 L}{2} \left\|\boldsymbol{X}\sqrt{\boldsymbol{Q}_{3\epsilon}^{(t)}}\right\|_2^2 \left\|\sqrt{\boldsymbol{Q}}\tilde{g}(\theta^{(t)})\right\|_2^2 \\
&\leq F(\theta^{(t)}) - \frac{\eta q^* \lambda^{\min}}{2} \left\|\sqrt{\boldsymbol{Q}}\tilde{g}(\theta^{(t)})\right\|_2^2 + \frac{\eta^2 L}{2} \|\boldsymbol{X}\|_2^2 (1+3\epsilon) \left\|\sqrt{\boldsymbol{Q}}\tilde{g}(\theta^{(t)})\right\|_2^2 \\
&\leq F(\theta^{(t)}) - \frac{\eta q^* \lambda^{\min}}{4} \left\|\sqrt{\boldsymbol{Q}}\tilde{g}(\theta^{(t)})\right\|_2^2 \\
&\leq F(\theta^{(t)}) - \frac{\eta q^{*2} \lambda^{\min}}{4} \left\|\tilde{g}(\theta^{(t)})\right\|_2^2
\end{aligned} \tag{136}$$

This shows that $F(\theta^{(t)})$ is monotonically non-increasing. Since $F(\theta) \geq 0$, $F(\theta^{(t)})$ must converge as $t \to \infty$, and we need to prove that it converges to 0. Suppose that $F(\theta^{(t)})$ does not converge to 0, then there exists a constant $C > 0$ such that $F(\theta^{(t)}) \geq 2C$ for all t. On the other hand, it is easy to see that there exists θ^* such that $\ell(\langle \theta^*, \boldsymbol{x}_i \rangle, y_i) < C$ for all i. (136) also implies that $\left\|\tilde{g}(\theta^{(t)})\right\|_2 \to 0$ as $t \to \infty$ because we must have $F(\theta^{(t)}) - F(\theta^{(t+1)}) \to 0$.

Note that from (134) we have

$$\left\|\theta^{(t+1)} - \theta^*\right\|_2^2 = \left\|\theta^{(t)} - \theta^*\right\|_2^2 + 2\eta \langle \boldsymbol{X}\boldsymbol{Q}^{(t)}\tilde{g}(\theta^{(t)}), \theta^* - \theta^{(t)} \rangle + \eta^2 \left\|\boldsymbol{X}\boldsymbol{Q}^{(t)}\tilde{g}(\theta^{(t)})\right\|_2^2 \tag{137}$$

Denote

$$F_t(\theta) = \sum_{i=1}^{n} q_i^{(t)} \ell(\langle \theta, \boldsymbol{x}_i \rangle, y_i) \tag{138}$$

Then F_t is convex because ℓ is convex and $q_i^{(t)}$ are non-negative, and $\nabla F_t(\theta^{(t)}) = \boldsymbol{X}\boldsymbol{Q}^{(t)}\tilde{g}(\theta^{(t)})$. By the lower linear bound $F_t(\boldsymbol{y}) \geq F_t(\boldsymbol{x}) + \langle \nabla F_t(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle$, we have for all t,

$$\langle \boldsymbol{X}\boldsymbol{Q}^{(t)}\tilde{g}(\theta^{(t)}), \theta^* - \theta^{(t)} \rangle \leq F_t(\theta^*) - F_t(\theta^{(t)}) \leq F_t(\theta^*) - \frac{2}{3}F(\theta^{(t)}) \leq C - \frac{4C}{3} = -\frac{C}{3} \tag{139}$$

because $q_i^{(t)} \geq q_i - \epsilon \geq \frac{2}{3}q_i$ and $\sum_{i=1}^n q_i^{(t)} = 1$. Since $\left\|\tilde{g}(\theta^{(t)})\right\|_2 \to 0$, there exists $T > 0$ such that for all $t \geq T$ and all $\eta \leq \eta_0$,

$$\left\|\theta^{(t+1)} - \theta^*\right\|_2^2 \leq \left\|\theta^{(t)} - \theta^*\right\|_2^2 - \frac{\eta C}{3} \tag{140}$$

which means that $\left\|\theta^{(t)} - \theta^*\right\|_2^2 \to -\infty$ because $\frac{\eta C}{3}$ is a positive constant. This is a contradiction! Thus, $F(\theta^{(t)})$ must converge to 0, which is result (i).

(i) immediately implies (ii) because ℓ is strictly decreasing to 0 by condition.

Now let's prove (iii). First of all, the uniqueness of θ_R can be easily proved from the convexity of $F(\theta)$. The condition implies that $y_i\langle \theta_R, \boldsymbol{x}_i \rangle > 0$, i.e. θ_R must classify all training samples correctly. If there are two different minimizers θ_R and θ_R' in whose norm is at most R, then consider $\theta_R'' = \frac{1}{2}(\theta_R + \theta_R')$. By the convexity of F, we know that θ_R'' must also be a minimizer, and $\|\theta_R''\|_2 < R$. Thus, $F(\frac{R}{\|\theta_R''\|_2}\theta_R'') < F(\theta_R'')$ and $\|\frac{R}{\|\theta_R''\|_2}\theta_R''\|_2 = R$, which contradicts with the fact that θ_R'' is a minimizer.

To prove the rest of (iii), the key is to consider (135). On one hand, similar to (36) we can prove that for all $t \geq t_\epsilon$, there is

$$\left|\langle \boldsymbol{X}\boldsymbol{Q}^{(t)}\tilde{g}(\theta^{(t)}), \boldsymbol{X}(\boldsymbol{Q}^{(t)} - \boldsymbol{Q})\tilde{g}(\theta^{(t)})\rangle\right| \leq \lambda^{\max}\sqrt{3\epsilon}\left\|\sqrt{\boldsymbol{Q}^{(t)}}\tilde{g}(\theta^{(t)})\right\|_2^2 \tag{141}$$

Since we choose $\epsilon = \min\{\frac{q^*}{3}, \frac{(q^*\lambda^{\min})^2}{192\lambda^{\max 2}}\}$, this inequality implies that

$$\begin{aligned}\left\|\nabla F_t(\theta^{(t)})\right\|_2^2 &= \left\|\boldsymbol{X}\boldsymbol{Q}^{(t)}\tilde{g}(\theta^{(t)})\right\|_2^2 \geq \lambda^{\min}\left\|\boldsymbol{Q}^{(t)}\tilde{g}(\theta^{(t)})\right\|_2^2 \geq \lambda^{\min}(q^* - \epsilon)\left\|\sqrt{\boldsymbol{Q}^{(t)}}\tilde{g}(\theta^{(t)})\right\|_2^2 \\ &\geq \frac{\lambda^{\min}q^*}{2}\left\|\sqrt{\boldsymbol{Q}^{(t)}}\tilde{g}(\theta^{(t)})\right\|_2^2 \geq 4\left|\langle \boldsymbol{X}\boldsymbol{Q}^{(t)}\tilde{g}(\theta^{(t)}), \boldsymbol{X}(\boldsymbol{Q}^{(t)} - \boldsymbol{Q})\tilde{g}(\theta^{(t)})\rangle\right|\end{aligned} \tag{142}$$

On the other hand, if $\eta \leq \eta_2 = \frac{1}{2L}$, we will have

$$\frac{\eta^2 L}{2}\left\|\boldsymbol{X}\boldsymbol{Q}^{(t)}\tilde{g}(\theta^{(t)})\right\|_2^2 \leq \frac{\eta}{4}\left\|\nabla F_t(\theta^{(t)})\right\|_2^2 \tag{143}$$

Combining all the above with (135) yields

$$F(\theta^{(t+1)}) - F(\theta^{(t)}) \leq -\frac{\eta}{2}\left\|\nabla F_t(\theta^{(t)})\right\|_2^2 \tag{144}$$

Denote $\boldsymbol{u} = \lim_{R\to\infty}\frac{\theta_R}{\|\theta_R\|_2}$. Similar to Lemma 9 in [JDST20], we can prove that: for any $\alpha > 0$, there exists a constant $\rho(\alpha) > 0$ such that for any θ subject to $\|\theta\|_2 \geq \rho(\alpha)$, there is

$$F_t((1+\alpha)\|\theta\|_2\boldsymbol{u}) \leq F_t(\theta) \tag{145}$$

for any t. Let $t_\alpha \geq t_\epsilon$ satisfy that for all $t \geq t_\alpha$, $\|\theta^{(t)}\|_2 \geq \max\{\rho(\alpha), 1\}$. By the convexity of F_t, for all $t \geq t_\alpha$,

$$\langle \nabla F_t(\theta^{(t)}), \theta^{(t)} - (1+\alpha)\|\theta^{(t)}\|_2\boldsymbol{u}\rangle \geq F_t(\theta^{(t)}) - F_t((1+\alpha)\|\theta^{(t)}\|_2\boldsymbol{u}) \geq 0 \tag{146}$$

Thus, we have

$$\begin{aligned}\langle \theta^{(t+1)} - \theta^{(t)}, \boldsymbol{u} \rangle &= \langle -\eta \nabla F_t(\theta^{(t)}), \boldsymbol{u} \rangle \\ &\geq \langle -\eta \nabla F_t(\theta^{(t)}), \theta^{(t)} \rangle \frac{1}{(1+\alpha)\|\theta^{(t)}\|_2} \\ &= \langle \theta^{(t+1)} - \theta^{(t)}, \theta^{(t)} \rangle \frac{1}{(1+\alpha)\|\theta^{(t)}\|_2} \\ &= \left(\frac{1}{2} \left\| \theta^{(t+1)} \right\|_2^2 - \frac{1}{2} \left\| \theta^{(t)} \right\|_2^2 - \frac{1}{2} \left\| \theta^{(t+1)} - \theta^{(t)} \right\|_2^2 \right) \frac{1}{(1+\alpha)\|\theta^{(t)}\|_2}\end{aligned} \tag{147}$$

By $\frac{1}{2}(\|\theta^{(t+1)}\|_2 - \|\theta^{(t)}\|_2)^2 \geq 0$, we have $(\frac{1}{2} \left\| \theta^{(t+1)} \right\|_2^2 - \frac{1}{2} \left\| \theta^{(t)} \right\|_2^2)/\|\theta^{(t)}\|_2 \geq \left\| \theta^{(t+1)} \right\|_2 - \left\| \theta^{(t)} \right\|_2$. Moreover, by (144) we have

$$\frac{\left\| \theta^{(t+1)} - \theta^{(t)} \right\|_2^2}{2(1+\alpha)\|\theta^{(t)}\|_2} \leq \frac{\left\| \theta^{(t+1)} - \theta^{(t)} \right\|_2^2}{2} = \frac{\eta^2 \left\| \nabla F_t(\theta^{(t)}) \right\|_2^2}{2} \leq \eta \left(F(\theta^{(t)}) - F(\theta^{(t+1)}) \right) \tag{148}$$

Summing up (147) from $t = t_\alpha$ to $t-1$, we have

$$\langle \theta^{(t)} - \theta^{(t_\alpha)}, \boldsymbol{u} \rangle \geq \frac{\left\| \theta^{(t)} \right\|_2 - \left\| \theta^{(t_\alpha)} \right\|_2}{1+\alpha} + \eta \left(F(\theta^{(t)}) - F(\theta^{(t_\alpha)}) \right) \geq \frac{\left\| \theta^{(t)} \right\|_2 - \left\| \theta^{(t_\alpha)} \right\|_2}{1+\alpha} - \eta F(\theta^{(t_\alpha)}) \tag{149}$$

which implies that

$$\left\langle \frac{\theta^{(t)}}{\left\| \theta^{(t)} \right\|_2}, \boldsymbol{u} \right\rangle \geq \frac{1}{1+\alpha} + \frac{1}{\left\| \theta^{(t)} \right\|_2} \left(\langle \theta^{(t_\alpha)}, \boldsymbol{u} \rangle - \frac{\|\theta^{(t_\alpha)}\|_2}{1+\alpha} - \eta F(\theta^{(t_\alpha)}) \right) \tag{150}$$

Since $\lim_{t\to\infty} \|\theta^{(t)}\|_2 = \infty$, we have

$$\liminf_{t\to\infty} \left\langle \frac{\theta^{(t)}}{\left\| \theta^{(t)} \right\|_2}, \boldsymbol{u} \right\rangle \geq \frac{1}{1+\alpha} \tag{151}$$

Since α is arbitrary, we must have $\lim_{t\to\infty} \frac{\theta^{(t)}}{\left\| \theta^{(t)} \right\|_2} = \boldsymbol{u}$ as long as $\eta \leq \min\{\eta_1, \eta_2\}$. □

D.5.3 Corollary of Theorem 8

We can show that for the logistic loss, it satisfies all conditions of Theorem 8 and $\lim_{R\to\infty} \frac{\theta_R}{R} = \hat{\theta}_{\text{MM}}$. First of all, for the logistic loss we have $\nabla_{\hat{y}}^2 \ell(\hat{y}, y) = \frac{y^2}{e^{y\hat{y}} + e^{-y\hat{y}} + 2} \leq \max_i \frac{y_i^2}{4}$, so ℓ is smooth.

Then, we prove that $\lim_{R\to\infty} \frac{\theta_R}{R}$ exists and is equal to $\hat{\theta}_{\text{MM}}$. For the logistic loss, it is easy to show that for any $\hat{\theta}' \neq \hat{\theta}_{\text{MM}}$, there exists an $R(\hat{\theta}') > 0$ and an $\delta(\hat{\theta}') > 0$ such that $F(R \cdot \theta) > F(R \cdot \hat{\theta}_{\text{MM}})$ for all $R \geq R(\hat{\theta}')$ and $\theta \in B(\hat{\theta}', \delta(\hat{\theta}'))$.

Let $S = \{\theta : \|\theta\|_2 = 1\}$. For any $\epsilon > 0$, $S - B(\hat{\theta}_{\text{MM}}, \epsilon)$ is a compact set. And for any $\theta \in S - B(\hat{\theta}_{\text{MM}}, \epsilon)$, there exist $R(\theta)$ and $\delta(\theta)$ as defined above. Thus, there must exist $\theta_1, \cdots, \theta_m \in S - B(\hat{\theta}_{\text{MM}}, \epsilon)$ such that $S - B(\hat{\theta}_{\text{MM}}, \epsilon) \subseteq \cup_{i=1}^m B(\theta_i, \delta(\theta_i))$. Let $R(\epsilon) = \max\{R(\theta_1), \cdots, R(\theta_m)\}$, then for all $R \geq R(\epsilon)$ and all $\theta \in S - B(\hat{\theta}_{\text{MM}}, \epsilon)$, $F(R \cdot \theta) > F(R \cdot \hat{\theta}_{\text{MM}})$, which means that $\frac{\theta_R}{R} \in B(\hat{\theta}_{\text{MM}}, \epsilon)$ for all $R \geq R(\epsilon)$. Therefore, $\lim_{R\to\infty} \frac{\theta_R}{R}$ exists and is equal to $\hat{\theta}_{\text{MM}}$.

Therefore, by Theorem 8, any GRW satisfying Assumption 1 makes a linear model converge to the max-margin classifier under the logistic loss.

D.6 Proof of Theorem 9

We first consider the regularized linearized neural network $f_{\text{linreg}}^{(t)}$. Since by Proposition 11 $f^{(0)}(\boldsymbol{x})$ is sampled from a zero-mean Gaussian process, there exists a constant $M > 0$ such that $|f^{(0)}(\boldsymbol{x}_i)| < M$

for all i with high probability. Define

$$F(\theta) = \sum_{i=1}^{n} q_i \ell(\langle \theta, \nabla_\theta f^{(0)}(\boldsymbol{x}_i)\rangle + f^{(0)}(\boldsymbol{x}_i), y_i) \tag{152}$$

Denote $\tilde{\theta}_R = \arg\min_\theta \{F(R \cdot \theta) : \|\theta\|_2 \leq 1\}$. when the linearized neural network is trained by a GRW satisfying Assumption 1 with regularization, since this is convex optimization and the objective function is smooth, we can prove that with a sufficiently small learning rate, as $t \to \infty$, $\theta^{(t)} \to R \cdot \tilde{\theta_R} + \theta^{(0)}$ where $R = \lim_{t\to\infty} \|\theta^{(t)} - \theta^{(0)}\|_2$ (which is the minimizer). And define

$$\gamma = \min_{i=1,\cdots,n} y_i \cdot \langle \hat{\theta}_{\text{MM}}, \nabla_\theta f^{(0)}(\boldsymbol{x}_i)\rangle \tag{153}$$

First, we derive the lower bound of R. By Theorem 16, with a sufficiently large $\tilde{d}$, with high probability $\hat{\mathcal{R}}(f^{(t)}_{\text{reg}}) < \epsilon$ implies $\hat{\mathcal{R}}(f^{(t)}_{\text{linreg}}) < 2\epsilon$. By the convexity of ℓ, we have

$$\begin{aligned} 2\epsilon &> \frac{1}{n}\sum_{i=1}^{n} \ell(\langle R\tilde{\theta}_R, \boldsymbol{x}_i\rangle + f^{(0)}(\boldsymbol{x}_i), y_i) \geq \log\left(1 + \exp\left(-\frac{1}{n}\sum_{i=1}^{n}(\langle R\tilde{\theta}_R, \boldsymbol{x}_i\rangle + f^{(0)}(\boldsymbol{x}_i))y_i\right)\right) \\ &\geq \log\left(1 + \exp\left(-\frac{1}{n}\sum_{i=1}^{n} R\langle \tilde{\theta}_R, \boldsymbol{x}_i\rangle y_i - M\right)\right) \end{aligned} \tag{154}$$

which implies that $R = \Omega(-\log 2\epsilon)$ for all $\epsilon \in (0, \frac{1}{4})$.

Denote $\delta = \|\hat{\theta}_{\text{MM}} - \tilde{\theta}_R\|_2$. Let $\theta' = \frac{\hat{\theta}_{\text{MM}} + \tilde{\theta}_R}{2}$, then we can see that $\|\theta'\|_2 = \sqrt{1 - \frac{\delta^2}{4}}$. Let $\tilde{\theta}' = \frac{\theta'}{\|\theta'\|_2}$. By the definition of $\hat{\theta}_{\text{MM}}$, there exists j such that $y_j \cdot \langle \tilde{\theta}', \nabla_\theta f^{(0)}(\boldsymbol{x}_j)\rangle \leq \gamma$, which implies

$$y_j \cdot \left\langle \frac{\hat{\theta}_{\text{MM}} + \tilde{\theta}_R}{2} \frac{1}{\sqrt{1 - \frac{\delta^2}{4}}}, \nabla_\theta f^{(0)}(\boldsymbol{x}_j) \right\rangle \leq \gamma \tag{155}$$

Thus, we have

$$\begin{aligned} y_j \cdot \langle \tilde{\theta}_R, \nabla_\theta f^{(0)}(\boldsymbol{x}_j)\rangle &\leq 2\sqrt{1 - \frac{\delta^2}{4}}\gamma - y_j \cdot \langle \hat{\theta}_{\text{MM}}, \nabla_\theta f^{(0)}(\boldsymbol{x}_j)\rangle \\ &\leq \left(2\sqrt{1 - \frac{\delta^2}{4}} - 1\right)\gamma \\ &\leq \left(2(1 - \frac{\delta^2}{8}) - 1\right)\gamma \qquad (\text{since } \sqrt{1-x} \leq 1 - \frac{x}{2}) \\ &= (1 - \frac{\delta^2}{4})\gamma \end{aligned} \tag{156}$$

On the other hand, we have

$$\begin{aligned} &q_j \log(1 + \exp(-y_j \cdot \langle R \cdot \tilde{\theta}_R, \nabla_\theta f^{(0)}(\boldsymbol{x}_j)\rangle - M)) \leq F(R \cdot \tilde{\theta}_R) \\ \leq &F(R \cdot \hat{\theta}_{\text{MM}}) \leq \log(1 + \exp(-R\gamma + M)) \end{aligned} \tag{157}$$

which implies that

$$q^* \log\left(1 + \exp\left(-(1 - \frac{\delta^2}{4})R\gamma - M\right)\right) \leq \log(1 + \exp(-R\gamma + M)) \tag{158}$$

and this leads to

$$1 + \exp(-R\gamma + M) \geq \left(1 + \exp\left(-(1 - \frac{\delta^2}{4})R\gamma - M\right)\right)^{q^*} \geq 1 + q^* \exp\left(-(1 - \frac{\delta^2}{4})R\gamma - M\right) \tag{159}$$

which is equivalent to

$$-R\gamma + M \geq -(1 - \frac{\delta^2}{4})R\gamma - M + \log(q^*) \tag{160}$$

Thus, we have

$$\delta = O(R^{-1/2}) = O((-\log 2\epsilon)^{-1/2}) \tag{161}$$

So for any test point $\boldsymbol{x}$, since $\|\nabla_\theta f^{(0)}(\boldsymbol{x})\|_2 \leq M_0$, we have

$$|\langle \hat{\theta}_{\text{MM}} - \tilde{\theta}_R, \nabla_\theta f^{(0)}(\boldsymbol{x})\rangle| \leq \delta M_0 = O((-\log 2\epsilon)^{-1/2}) \tag{162}$$

Combined with Theorem 16, we have: with high probability,

$$\limsup_{t\to\infty} |R \cdot f_{\text{MM}}(\boldsymbol{x}) - f_{\text{reg}}^{(t)}(\boldsymbol{x})| = O(R \cdot (-\log 2\epsilon)^{-1/2} + \tilde{d}^{-1/4}) \tag{163}$$

So there exists a constant $C > 0$ such that: As $\tilde{d} \to \infty$, with high probability, for all $\epsilon \in (0, \frac{1}{4})$, if $|f_{\text{MM}}(\boldsymbol{x})| > C \cdot (-\log 2\epsilon)^{-1/2}$, then $f_{\text{reg}}^{(t)}(\boldsymbol{x})$ will have the same sign as $f_{\text{MM}}(\boldsymbol{x})$ for a sufficiently large t. Note that this C only depends on n, q^*, γ, M and M_0, so it is a constant independent of ϵ. □

Remark. Note that Theorem 9 requires Assumption 1 while Theorem 6 does not due to the fundamental difference between the classification and regression. In regression the model converges to a finite point. However, in classification, the training loss converging to zero implies that either (i) The direction of the weight is close to the max-margin classifier or (ii) The norm of the weight is very large. Assumption 1 is used to eliminate the possibility of (ii). If the regularization parameter μ is sufficiently large, then a small empirical risk could imply a small weight norm. However, in our theorem we do not assume anything on μ, so Assumption 1 is necessary.

E A Note on the Proofs in [LXS$^+$19]

We have mentioned that the proofs in [LXS$^+$19], particularly the proofs of their Theorem 2.1 and Lemma 1 in their Appendix G, are flawed. In order to fix their proof, we change the network initialization to (9). In this section, we will demonstrate what goes wrong in the proofs in [LXS$^+$19], and how we manage to fix the proof. For clarity, we are referring to the following version of the paper: `https://arxiv.org/pdf/1902.06720v4.pdf`.

To avoid confusion, in this section we will still use the notations used in our paper.

E.1 Their Problems

[LXS$^+$19] claimed in their Theorem 2.1 that under the conditions of our Lemma 5, for any $\delta > 0$, there exist $\tilde{D} > 0$ and a constant C such that for any $\tilde{d} \geq \tilde{D}$, with probability at least $(1 - \delta)$, the gap between the output of a sufficiently wide fully-connected neural network and the output of its linearized neural network at any test point $\boldsymbol{x}$ can be uniformly bounded by

$$\sup_{t\geq 0} \left| f^{(t)}(\boldsymbol{x}) - f_{\text{lin}}^{(t)}(\boldsymbol{x}) \right| \leq C\tilde{d}^{-1/2} \qquad \text{(claimed)} \tag{164}$$

where they used the original NTK formulation and initialization in [JGH18]:

$$\begin{cases} \boldsymbol{h}^{l+1} = \frac{W^l}{\sqrt{d_l}}\boldsymbol{x}^l + \beta \boldsymbol{b}^l \\ \boldsymbol{x}^{l+1} = \sigma(\boldsymbol{h}^{l+1}) \end{cases} \quad \text{and} \quad \begin{cases} W_{i,j}^{l(0)} \sim \mathcal{N}(0,1) \\ b_i^{l(0)} \sim \mathcal{N}(0,1) \end{cases} \quad (\forall l = 0, \cdots, L) \tag{165}$$

where $\boldsymbol{x}_0 = \boldsymbol{x}$ and $f(\boldsymbol{x}) = h^{L+1}$. However, in their proof in their Appendix G, they did not directly prove their result for the NTK formulation, but instead they proved another result for the following formulation which they called the *standard formulation*:

$$\begin{cases} \boldsymbol{h}^{l+1} = W^l \boldsymbol{x}^l + \beta \boldsymbol{b}^l \\ \boldsymbol{x}^{l+1} = \sigma(\boldsymbol{h}^{l+1}) \end{cases} \quad \text{and} \quad \begin{cases} W_{i,j}^{l(0)} \sim \mathcal{N}(0, \frac{1}{d_l}) \\ b_i^{l(0)} \sim \mathcal{N}(0,1) \end{cases} \quad (\forall l = 0, \cdots, L) \tag{166}$$

See their Appendix F for the definition of their standard formulation. In the original formulation, they also included two constants σ_w and σ_b for standard deviations, and for simplicity we omit these constants here. Note that the outputs of the NTK formulation and the standard formulation at initialization are actually the same. The only difference is that the norm of the weight W^l and the gradient of the model output with respect to W^l are different for all l.

In their Appendix G, they claimed that if a network with the standard formulation is trained by minimizing the squared loss with gradient descent and learning rate $\eta' = \eta/\tilde{d}$, where η is our learning rate in Lemma 5 and also their learning rate in their Theorem 2.1, then (164) is true for this network, so it is also true for a network with the NTK formulation because the two formulations have the same network output. And then they claimed in their equation (S37) that applying learning rate η' to the standard formulation is equivalent to applying the following learning rates

$$\eta_W^l = \frac{d_l}{d_{\max}}\eta \qquad \text{and} \qquad \eta_{\boldsymbol{b}}^l = \frac{1}{d_{\max}}\eta \tag{167}$$

to W^l and $\boldsymbol{b}^l$ of the NTK formulation, where $d_{\max} = \max\{d_0, \cdots, d_L\}$.

To avoid confusion, in the following discussions we will still use the NTK formulation and initialization if not stated otherwise.

Problem 1. Claim (167) is true, but it leads to two problems. The first problem is that $\eta_{\boldsymbol{b}}^l = O(d_{\max}^{-1})$ since $\eta = O(1)$, while their Theorem 2.1 needs the learning rate to be $O(1)$. Nevertheless, this problem can be simply fixed by modifying their standard formulation as $\boldsymbol{h}^{l+1} = W^l\boldsymbol{x}^l + \beta\sqrt{d_l}\boldsymbol{b}^l$ where $b_i^{l(0)} \sim \mathcal{N}(0, d_l^{-1})$. The real problem that is non-trivial to fix is that by (167), there is $\eta_W^0 = \frac{d_0}{d_{\max}}\eta$. However, note that d_0 is a constant since it is the dimension of the input space, while $d_{\max}$ goes to infinity. Consequently, in (167) they were essentially using a very small learning rate for the first layer W^0 but a normal learning rate for the rest of the layers, which definitely does not match with their claim in their Theorem 2.1.

Problem 2. Another big problem is that the proof of their Lemma 1 in their Appendix G is erroneous, and consequently their Theorem 2.1 is unsound as it heavily depends on their Lemma 1. In their Lemma 1, they claimed that for some constant $M > 0$, for any two models with the parameters θ and $\tilde{\theta}$ such that $\theta, \tilde{\theta} \in B(\theta^{(0)}, C_0)$ for some constant C_0, there is

$$\left\|J(\theta) - J(\tilde{\theta})\right\|_F \leq \frac{M}{\sqrt{\tilde{d}}}\left\|\theta - \tilde{\theta}\right\|_2 \qquad \text{(claimed)} \tag{168}$$

Note that the original claim in their paper was $\left\|J(\theta) - J(\tilde{\theta})\right\|_F \leq M\sqrt{\tilde{d}}\left\|\theta - \tilde{\theta}\right\|_2$. This is because they were proving this result for their standard formulation. Compared to the standard formulation, in the NTK formulation θ is $\sqrt{\tilde{d}}$ times larger, while the Jacobian $J(\theta)$ is $\sqrt{\tilde{d}}$ times smaller. This is also why here we have $\theta, \tilde{\theta} \in B(\theta^{(0)}, C_0)$ instead of $\theta, \tilde{\theta} \in B(\theta^{(0)}, C_0\tilde{d}^{-1/2})$ for the NTK formulation. Therefore, equivalently they were claiming (168) for the NTK formulation.

However, their proof of (168) in incorrect. Specifically, the right-hand side of their inequality (S86) is incorrect. Using the notations in our Appendix D.3.3, their (S86) essentially claimed that

$$\left\|\boldsymbol{\alpha}^l - \tilde{\boldsymbol{\alpha}}^l\right\|_2 \leq \frac{M}{\sqrt{\tilde{d}}}\left\|\theta - \tilde{\theta}\right\|_2 \qquad \text{(claimed)} \tag{169}$$

for any $\theta, \tilde{\theta} \in B(\theta^{(0)}, C_0)$, where $\boldsymbol{\alpha}^l = \nabla_{\boldsymbol{h}^l}\boldsymbol{h}^{L+1}$ and $\tilde{\boldsymbol{\alpha}}^l$ is the same gradient for the second model. Note that their (S86) does not have the $\sqrt{\tilde{d}}$ in the denominator which appears in (169). This is because for their standard formulation, θ is $\sqrt{\tilde{d}}$ times smaller than the original NTK formulation, while $\left\|\boldsymbol{\alpha}^l\right\|_2$ has the same order in the two formulations because all $\boldsymbol{h}^l$ are the same.

However, it is actually impossible to prove (169). Consider the following counterexample: Since θ and $\tilde{\theta}$ are arbitrarily chosen, we can choose them such that they only differ in b_1^l for some $1 \leq l < L$. Then, $\left\|\theta - \tilde{\theta}\right\|_2 = \left|b_1^l - \tilde{b}_1^l\right|$. We can see that $\boldsymbol{h}^{l+1}$ and $\tilde{\boldsymbol{h}}^{l+1}$ only differ in the first element, and

$\left|h_1^{l+1} - \tilde{h}_1^{l+1}\right| = \left|\beta(b_1^l - \tilde{b}_1^l)\right|$. Moreover, we have $W^{l+1} = \tilde{W}^{l+1}$, so there is

$$\begin{aligned}\boldsymbol{\alpha}^{l+1} - \tilde{\boldsymbol{\alpha}}^{l+1} =& \text{diag}(\dot{\sigma}(\boldsymbol{h}^{l+1}))\frac{W^{l+1\top}}{\sqrt{\tilde{d}}}\boldsymbol{\alpha}^{l+2} - \text{diag}(\dot{\sigma}(\tilde{\boldsymbol{h}}^{l+1}))\frac{\tilde{W}^{l+1\top}}{\sqrt{\tilde{d}}}\tilde{\boldsymbol{\alpha}}^{l+2} \\ =& \left[\text{diag}(\dot{\sigma}(\boldsymbol{h}^{l+1})) - \text{diag}(\dot{\sigma}(\tilde{\boldsymbol{h}}^{l+1}))\right]\frac{W^{l+1\top}}{\sqrt{\tilde{d}}}\boldsymbol{\alpha}^{l+2} \\ &+ \text{diag}(\dot{\sigma}(\tilde{\boldsymbol{h}}^{l+1}))\frac{W^{l+1\top}}{\sqrt{\tilde{d}}}(\boldsymbol{\alpha}^{l+2} - \tilde{\boldsymbol{\alpha}}^{l+2})\end{aligned} \tag{170}$$

Then we can lower bound $\left\|\boldsymbol{\alpha}^{l+1} - \tilde{\boldsymbol{\alpha}}^{l+1}\right\|_2$ by

$$\begin{aligned}\left\|\boldsymbol{\alpha}^{l+1} - \tilde{\boldsymbol{\alpha}}^{l+1}\right\|_2 \geq& \left\|\left[\text{diag}(\dot{\sigma}(\boldsymbol{h}^{l+1})) - \text{diag}(\dot{\sigma}(\tilde{\boldsymbol{h}}^{l+1}))\right]\frac{W^{l+1\top}}{\sqrt{\tilde{d}}}\boldsymbol{\alpha}^{l+2}\right\|_2 \\ &- \left\|\text{diag}(\dot{\sigma}(\tilde{\boldsymbol{h}}^{l+1}))\frac{W^{l+1\top}}{\sqrt{\tilde{d}}}(\boldsymbol{\alpha}^{l+2} - \tilde{\boldsymbol{\alpha}}^{l+2})\right\|_2\end{aligned} \tag{171}$$

The first term on the right-hand side is equal to $\left|\left[\dot{\sigma}(h_1^{l+1}) - \dot{\sigma}(\tilde{h}_1^{l+1})\right]\langle W_1^{l+1}/\sqrt{\tilde{d}}, \boldsymbol{\alpha}^{l+2}\rangle\right|$ where W_1^{l+1} is the first row of W^{l+1}. We know that $\left\|W_1^{l+1}\right\|_2 = \Theta\left(\sqrt{\tilde{d}}\right)$ with high probability as its elements are sampled from $\mathcal{N}(0,1)$, and in their (S85) they claimed that $\left\|\boldsymbol{\alpha}^{l+2}\right\|_2 = O(1)$, which is true. In addition, they assumed that $\dot{\sigma}$ is Lipschitz. Hence, we can see that

$$\left\|\left[\text{diag}(\dot{\sigma}(\boldsymbol{h}^{l+1})) - \text{diag}(\dot{\sigma}(\tilde{\boldsymbol{h}}^{l+1}))\right]\frac{W^{l+1\top}}{\sqrt{\tilde{d}}}\boldsymbol{\alpha}^{l+2}\right\|_2 = O\left(\left|h_1^{l+1} - \tilde{h}_1^{l+1}\right|\right) = O\left(\left\|\theta - \tilde{\theta}\right\|_2\right) \tag{172}$$

On the other hand, suppose that claim (169) is true, then $\left\|\boldsymbol{\alpha}^{l+2} - \tilde{\boldsymbol{\alpha}}^{l+2}\right\|_2 = O\left(\tilde{d}^{-1/2}\left\|\theta - \tilde{\theta}\right\|_2\right)$. Then we can see that the second term on the right-hand side is $O\left(\tilde{d}^{-1/2}\left\|\theta - \tilde{\theta}\right\|_2\right)$ because $\left\|W^{l+1}\right\|_2 = O(\sqrt{\tilde{d}})$ and $\dot{\sigma}(x)$ is bounded by a constant as σ is Lipschitz. Thus, for a very large $\tilde{d}$, the second-term is an infinitesimal compared to the first term, so we can only prove that

$$\left\|\boldsymbol{\alpha}^{l+1} - \tilde{\boldsymbol{\alpha}}^{l+1}\right\|_2 = O\left(\left\|\theta - \tilde{\theta}\right\|_2\right) \tag{173}$$

which is different from (169) because it lacks a critical $\tilde{d}^{-1/2}$ and thus leads to a contradiction. Hence, we cannot prove (169) with the $\tilde{d}^{-1/2}$ factor, and consequently we cannot prove (168) with the $\sqrt{\tilde{d}}$ in the denominator on the right-hand side. As a result, their Lemma 1 and Theorem 2.1 cannot be proved without this critical $\tilde{d}^{-1/2}$. Similarly, we can also construct a counterexample where θ and $\tilde{\theta}$ only differ in the first row of some W^l.

E.2 Our Fixes

Regarding Problem 1, we can still use an $O(1)$ learning rate for the first layer in the NTK formulation given that $\|\boldsymbol{x}\|_2 \leq 1$. This is because for the first layer, we have

$$\nabla_{W^0} f(\boldsymbol{x}) = \frac{1}{\sqrt{d_0}}\boldsymbol{x}^0\alpha^{1\top} = \frac{1}{\sqrt{d_0}}\boldsymbol{x}\alpha^{1\top} \tag{174}$$

For all $l \geq 1$, we have $\left\|\boldsymbol{x}^l\right\|_2 = O(\tilde{d}^{1/2})$. However, for $l = 0$, we instead have $\left\|\boldsymbol{x}^0\right\|_2 = O(1)$. Thus, we can prove that the norm of $\nabla_{W^0} f(\boldsymbol{x})$ has the same order as the gradient with respect to any other layer, so there is no need to use a smaller learning rate for the first layer.

Regarding Problem 2, in our formulation (8) and initialization (9), the initialization of the last layer of the NTK formulation is changed from the Gaussian initialization $W_{i,j}^{L(0)} \sim \mathcal{N}(0,1)$ to the zero initialization $W_{i,j}^{L(0)} = 0$. Now we show how this modification solves Problem 2.

The main consequence of changing the initialization of the last layer is that (86) becomes different: instead of $\left\|W^L\right\|_2 \leq 3\sqrt{\tilde{d}}$, we now have $\left\|W^L\right\|_2 \leq C_0 \leq 3\sqrt[4]{\tilde{d}}$. In fact, for any $r \in (0, 1/2)$, we can prove that $\left\|W^L\right\|_2 \leq 3\tilde{d}^r$ for sufficiently large $\tilde{d}$. In our proof we choose $r = 1/4$.

Consequently, instead of $\left\|\boldsymbol{\alpha}^l\right\|_2 \leq M_3$, we can now prove that $\left\|\boldsymbol{\alpha}^l\right\|_2 \leq M_3\tilde{d}^{r-1/2}$ for all $l \leq L$ by induction. So now we can prove $\left\|\boldsymbol{\alpha}^l - \tilde{\boldsymbol{\alpha}}^l\right\|_2 = O\left(\tilde{d}^{r-1/2}\left\|\theta - \tilde{\theta}\right\|_2\right)$ instead of $O\left(\left\|\theta - \tilde{\theta}\right\|_2\right)$, because

- For $l < L$, we now have $\left\|\boldsymbol{\alpha}^{l+1}\right\|_2 = O(\tilde{d}^{r-1/2})$ instead of $O(1)$, so we can have the additional $\tilde{d}^{r-1/2}$ factor in the bound.
- For $l = L$, although $\left\|\boldsymbol{\alpha}^{L+1}\right\|_2 = 1$, note that $\left\|W^L\right\|_2$ now becomes $O(\tilde{d}^r)$ instead of $O(\tilde{d}^{1/2})$, so again we can decrease the bound by a factor of $\tilde{d}^{r-1/2}$.

Then, with this critical $\tilde{d}^{r-1/2}$, we can prove the approximation theorem with the form

$$\sup_{t\geq 0}\left|f^{(t)}(\boldsymbol{x}) - f_{\text{lin}}^{(t)}(\boldsymbol{x})\right| \leq C\tilde{d}^{r-1/2} \tag{175}$$

for any $r \in (0, 1/2)$, though we cannot really prove the $O(\tilde{d}^{-1/2})$ bound as originally claimed in (164). So this is how we solve Problem 2.

One caveat of changing the initialization to zero initialization is whether we can still safely assume that $\lambda^{\min} > 0$ where $\lambda^{\min}$ is the smallest eigenvalue of Θ, the kernel matrix of our new formulation. The answer is yes. In fact, in our Proposition 3 we proved that Θ is non-degenerated (which means that $\Theta(\boldsymbol{x}, \boldsymbol{x}')$ still depends on $\boldsymbol{x}$ and $\boldsymbol{x}'$), and under the overparameterized setting where $d_L \gg n$, chances are high that Θ is full-rank. Hence, we can still assume that $\lambda^{\min} > 0$.

As a final remark, one key reason why we need to initialize W^L as zero is that the dimension of the output space (i.e. the dimension of $\boldsymbol{h}^{L+1}$) is finite, and in our case it is 1. Suppose we allow the dimension of $\boldsymbol{h}^{L+1}$ to be $\tilde{d}$ which goes to infinity, then using the same proof techniques, for the NTK formulation we can prove that $\sup_t \left\|\boldsymbol{h}^{L+1(t)} - \boldsymbol{h}_{\text{lin}}^{L+1(t)}\right\|_2 \leq C$, i.e. the gap between two vectors of infinite dimension is always bounded by a finite constant. This is the approximation theorem we need for the infinite-dimensional output space. However, when the dimension of the output space is finite, $\sup_t \left\|\boldsymbol{h}^{L+1(t)} - \boldsymbol{h}_{\text{lin}}^{L+1(t)}\right\|_2 \leq C$ no longer suffices, so we need to decrease the order of the norm of W^L in order to obtain a smaller upper bound.

References

[AZLS19] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pages 242–252. PMLR, 2019.

[COB19] Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[DHP+12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[DLL+19] Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pages 1675–1685. PMLR, 2019.

[DN18] John Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *arXiv preprint arXiv:1810.08750*, 2018.

[HPS16] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29, pages 3315–3323. Curran Associates, Inc., 2016.

[HSNL18] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In Jennifer Dy and Andreas Krause, editors, *International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1929–1938, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[JDST20] Ziwei Ji, Miroslav Dudík, Robert E. Schapire, and Matus Telgarsky. Gradient descent follows the regularization path for general losses. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2109–2136. PMLR, 09–12 Jul 2020.

[JGH18] Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[KLRS17] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in neural information processing systems*, pages 4066–4076, 2017.

[LHC+21] Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pages 6781–6792. PMLR, 2021.

[LXS+19] Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in neural information processing systems*, 32:8572–8583, 2019.

[OSHL19] Yonatan Oren, Shiori Sagawa, Tatsunori Hashimoto, and Percy Liang. Distributionally robust language modeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4227–4237, Hong Kong, China, November 2019. Association for Computational Linguistics.

[Raw01] John Rawls. *Justice as fairness: A restatement*. Harvard University Press, 2001.

[SKHL20] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020.

[SRKL20] Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8346–8356. PMLR, 13–18 Jul 2020.

[Ver10] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[ZDKR21] Runtian Zhai, Chen Dan, Zico Kolter, and Pradeep Ravikumar. Doro: Distributional and outlier robust optimization. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12345–12355. PMLR, 18–24 Jul 2021.

[ZVGRG17] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pages 1171–1180, 2017.

[ZWS+13] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, pages 325–333, 2013.