Blessing of Nonconvexity in Deep Linear Models: Depth Flattens the Optimization Landscape Around the True Solution

Jianhao Ma[*] and Salar Fattahi[+]

Department of Industrial and Operations Engineering
University of Michigan, Ann Arbor
[*]jianhao@umich.edu, [+]fattahi@umich.edu

July 15, 2022

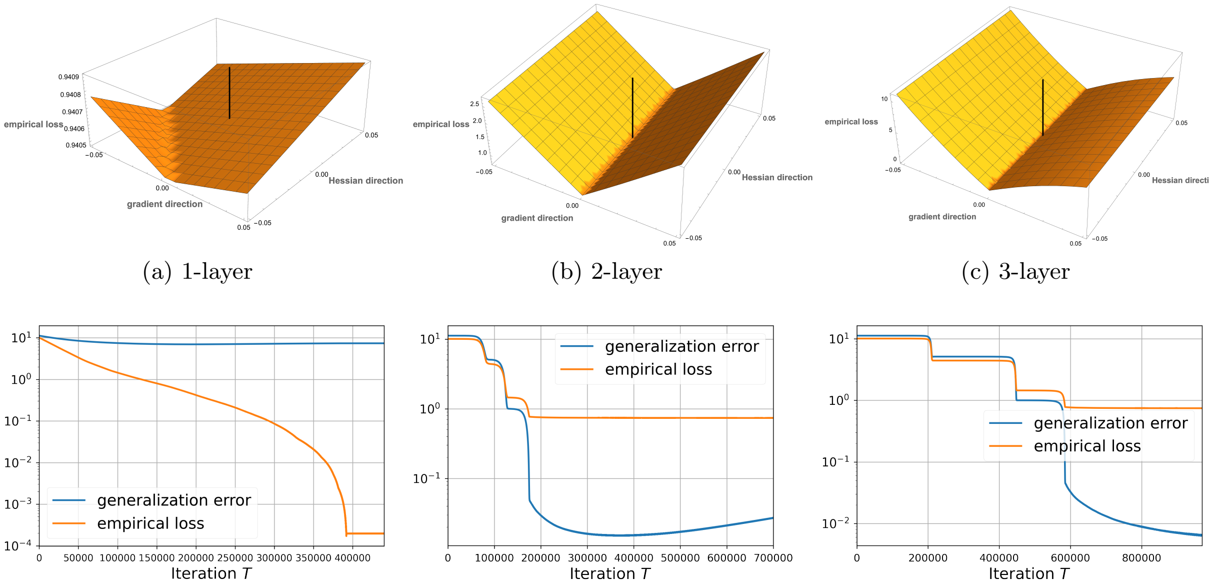
Abstract

This work characterizes the effect of depth on the optimization landscape of linear regression, showing that, despite their nonconvexity, deeper models have more desirable optimization landscape. We consider a robust and over-parameterized setting, where a subset of measurements are grossly corrupted with noise and the true linear model is captured via an N-layer linear neural network. On the negative side, we show that this problem *does not* have a benign landscape: given any $N \geq 1$, with constant probability, there exists a solution corresponding to the ground truth that is neither local nor global minimum. However, on the positive side, we prove that, for any N-layer model with $N \geq 2$, a simple sub-gradient method becomes oblivious to such "problematic" solutions; instead, it converges to a balanced solution that is not only close to the ground truth but also enjoys a flat local landscape, thereby eschewing the need for "early stopping". Lastly, we empirically verify that the desirable optimization landscape of deeper models extends to other robust learning tasks, including deep matrix recovery and deep ReLU networks with ℓ_1-loss.

1 Introduction

Supported by the empirical success of deep models in contemporary learning tasks, it is by now a conventional wisdom that "deeper models generalize better" [22, 32, 8]. Indeed, the flurry of recent attempts towards demystifying this phenomenon is a testament to the amount of research it has spawned: from simple linear regression to more complex and nonlinear models, it is shown that deeper models benefit from a range of desirable statistical properties, such as *depth separation* [34, 16, 35, 36], *benign overfitting* [4], and *hierarchical learning* [1], to name a few.

Despite the great promise of deeper models—both theoretically and empirically—the effect of depth on their optimization landscape has remained elusive to this day. A recent body of work attempts to characterize the effect of depth on the loss function through the notion of *benign landscape*. Roughly speaking, an optimization problem has a benign landscape if it is devoid of spurious local minima, and its true solutions—i.e., solutions corresponding to the ground truth—coincide with global minima. It has been shown that 2-layer [6] and multi-layer [24] linear neural networks with nearly-noiseless data have benign landscape. However, the notion of benign landscape is significantly



(a) 1-layer

(b) 2-layer

(c) 3-layer

(d) 1-layer

(e) 2-layer

(f) 3-layer

Figure 1: **First row.** Local landscape around the balanced true solution $\mathbf{w}^\star = (\sqrt[N]{\theta^\star}, \ldots, \sqrt[N]{\theta^\star})$ for 1-, 2-, and 3-layer models. x and y axis correspond to the points of the form $\mathbf{w}^\star + \alpha\mathbf{d} + \beta\mathbf{h}$, for different values of α and β, where $\mathbf{d}$ and $\mathbf{h}$ are respectively the most descent sub-gradient direction and the most negatively curved direction of the Hessian after smoothing. **Second row.** Generalization error and the empirical loss of the solutions found by SubGM for 1-, 2-, and 3-layer models.

stronger than what is needed in practice. For instance, the existence of spurious local minima may not pose any issue if an optimization algorithm can avoid them efficiently. Another line of research focuses on characterizing the solution trajectory of different local-search algorithms, showing that they enjoy an *implicit bias* that steers them away from undesirable solutions [26, 39, 11, 20, 2, 19]. However, such guarantees only apply to specific trajectories of an algorithm, thereby falling short of any meaningful characterization of the optimization landscape around those trajectories.

1.1 Our Contributions

To shed light on the effect of depth on the optimization landscape of deep models, we consider a prototypical problem in machine learning, namely *robust linear regression*, where the goal is to recover a linear model from a limited number of grossly corrupted measurements. Given samples of the form $y_i = \langle \theta^\star, x_i \rangle + \epsilon_i$, we study the optimization landscape of ℓ_1-loss under an N-layer model defined as $y = f_{\mathbf{w}}(x) := \langle w_1 \odot w_2 \odot \cdots \odot w_N, x \rangle$. Our results are summarized as follows:

- We prove that, for any $N \geq 1$, there exists at least one true solution that is neither local nor global minimum of ℓ_1-loss, provided that at least a fraction $p > 0$ of the measurements are corrupted with noise.
- Despite the ubiquity of such "hidden" true solutions, we show that, for any N-layer model with $N \geq 2$, a simple sub-gradient method (SubGM) with small initialization converges to a small neighborhood of a balanced true solution. The radius of this neighborhood shrinks

with the depth of the model, resulting in more accurate solutions. Moreover, the balancedness of the solution implies that each layer of the model inherits a similar sparsity pattern to the ground truth.

- We prove that deeper models take longer to train, but once trained, the algorithm will stay close to the ground truth for a longer time. This implies that early stopping of the algorithm becomes less crucial for deeper models.
- Finally, we prove that depth flattens the optimization landscape around the solution obtained by SubGM. In particular, we show that, within an γ-neighborhood of the true solution, the steepest descent direction can reduce the loss by at most $\mathcal{O}\left(\gamma^N\right)$, which decreases *exponentially* with N.

Motivating Example. To showcase our results, we consider an instance of robust linear regression in the over-parameterized setting, where the dimension of $\theta^\star$ is 500 and the number of available samples is 300. Moreover, we assume that 10% of the measurements are corrupted with large noise. The first row of Figure 1 shows the landscape around the balanced ground truth $w_1^\star = \sqrt[N]{\theta^\star}, w_2^\star = \sqrt[N]{\theta^\star}, \dots, w_N^\star = \sqrt[N]{\theta^\star}$.[1] In particular, x and y axis show the most descent sub-gradient direction and the most negatively curved direction of the Hessian after smoothing.[2] Evidently, there is a sharp transition in the landscape of N-layer models: for 1-layer model, the true solution has strictly negative directions along both sub-gradient and negative curvature. However, these descent directions almost disappear in 2- and 3-layer models. The second row of Figure 1 shows the performance of SubGM on these models. It can be observed that a 1-layer model easily overfits to noise, leading to a vacuous generalization error. On the contrary, a 3-layer model can find a solution that generalizes progressively better than 1- and 2-layer models, demonstrating the algorithmic benefit of the depth.

2 Related Works

Deep models: It is known that deeper models enjoy better *approximation power*. For instance, [16, 35, 36] introduce several functions that are expressible by deep models of moderate size; yet, they cannot be approximated via any shallow network of sub-exponential size. A recent work by [34] shows that depth separation may lead to optimization separation. In other words, functions that can be expressed by deeper models can also be efficiently learned via gradient descent. Another line of work shows that deep linear models have strong implicit bias towards the true solution [20, 2, 11], and that they benefit from *incremental learning* [19, 28]. More generally, [1] show that stochastic gradient descent on deep nonlinear models can provably learn certain complex functions by automatically decomposing them into a series of simpler ones.

Robust and sparse linear regression: Robust and sparse linear regression is a classical problem in statistics, with a wide range of applications in signal and image processing. Regularized methods, including Lasso [37, 10, 7, 33], Best Subset Selection [31, 5], and Forward and Backward Step-wise Regression [18], are considered as most widely used methods for solving robust and sparse linear

[1] Later, we will show that a simple sub-gradient method converges to this balanced solution.

[2] To smoothen $|x|$, we replace it with $\sqrt{x^2+\epsilon}$, for $\epsilon = 10^{-7}$.

regression that come equipped with strong statistical and computational guarantees. Recently, sparse linear regression has been used to explore the implicit bias of different optimization algorithms and initialization regimes. [39, 43] show that, for some unregularized overparameterized models, gradient descent (with early stopping) for sparse linear regression achieves minimax error rate. Moreover, [42] study how the scale of the initial point controls the transition between the "kernel" (lazy training) and "rich" regimes, and their corresponding generalization performance. [21] use this problem setting to explore the role of label noise in stochastic gradient descent.

Notations: For two vectors $x, y \in \mathbf{R}^d$, their inner product is defined as $\langle x, y\rangle = x^\top y$, and their Hadamard product is defined as $x \odot y = [x_1 y_1 \ \cdots \ x_d y_d]^\top$. For simplicity of notation, we use $\prod_j w_j$ to denote the Hadamard product of $w_1, w_2, \dots, w_N \in \mathbf{R}^d$. For a vector x, $\|x\|, \|x\|_\infty$, and $\|x\|_0$ refer to 2-norm, ∞-norm, and the number of nonzero elements, respectively. The symbols $a_t \lesssim b_t$ and $a_t = \mathcal{O}(b_t)$ are used to denote $a_t \le C b_t$, for a universal constant C. The notation $a_t = \Theta(b_t)$ is used to denote $a_t = O(b_t)$ and $b_t = \Omega(a_t)$. The $\mathrm{Sign}(\cdot)$ function is defined as $\mathrm{Sign}(x) = x/|x|$ if $x \neq 0$, and $\mathrm{Sign}(0) = [-1, 1]$. We denote $[n] := \{1, 2, \cdots, n\}$. For a vector $x \in \mathbb{R}^d$, we define $x^a = [x_1^a \ x_2^a \ \dots \ x_d^a]^\top$, for any $a > 0$. In all of our probabilistic arguments, the randomness is only over the input data and noise.

3 Problem Formulation

We study the problem of robust and sparse linear regression, where the goal is to estimate a k-sparse vector $\theta^\star \in \mathbf{R}^d$ ($k \ll d$) from a limited number of data points $\{(x_i, y_i)\}_{i=1}^m$, where $y_i = \langle \theta^\star, x_i\rangle + \epsilon_i$, x_i is i.i.d. standard Gaussian vector, and ϵ_i is noise. Moreover, for simplicity of our subsequent analysis, we assume that $\theta^\star$ is a non-negative vector. We believe that this assumption can be relaxed without a significant change in our results.

Assumption 1 (Noise Model)**.** *Given a corruption probability p, the noise vector $\epsilon = [\epsilon_1 \ \cdots \ \epsilon_m]^\top \in \mathbf{R}^m$ is generated as follows: first, a subset $\mathcal{S} \subset [m]$ with cardinality pm is chosen uniformly at random*[3]*. Then, for each entry $i \in \mathcal{S}$, the value of ϵ_i is drawn independently from a distribution P_o, and all the remaining entries are set to zero. Moreover, a random variable ζ under the distribution P_o satisfies $\mathbb{E}_{P_o}[\zeta] = 0$ and $\mathbb{P}(|\zeta| \ge t_0) \ge p_0$, for some strictly positive constants t_0 and p_0.*

Our considered noise model does not impose any assumption on the magnitude of the noise or the specific form of its distribution, which makes it particularly suitable for modeling outliers. Note that the assumption $\mathbb{P}(|\epsilon| \ge t_0) \ge p_0$ is very mild and satisfied for almost all common distributions. Roughly speaking, it implies that the noise takes a nonzero value with a nonzero probability.

To capture the input-output relationship, we consider a class of N-layer diagonal linear neural networks of the form $f_{\mathbf{w}}(x) = \langle w_1 \odot \cdots \odot w_N, x\rangle$, where $\mathbf{w} := (w_1, \cdots, w_N)$ collects the weights of the layers $w_1, \cdots, w_N \in \mathbf{R}^d$. Due to the sparse-and-large nature of the noise, it is natural to minimize the so-called empirical risk with ℓ_1-loss:

$$\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}) := \frac{1}{m}\sum_{i=1}^m |f_{\mathbf{w}}(x_i) - y_i| = \frac{1}{m}\sum_{i=1}^m |\langle w_1 \odot \cdots \odot w_N, x_i\rangle - y_i| \,. \tag{1}$$

[3] Here, for simplicity we assume pm is an integer.

Other variants of empirical risk minimization for linear regression have been studied in the literature. For instance, [11] study the solution trajectory of gradient flow on ℓ_2-loss, showing that it converges to a solution with smallest ℓ_1-norm. Similar analysis has also been appeared in more general deep linear neural networks [14, 15]. However, it is well-known that ℓ_2-loss is highly sensitive to outliers, and ℓ_1-loss is a better alternative to identify and reject large-and-sparse noise.

A solution $\bar{\mathbf{w}}$ is called *global* if it corresponds to a global minimizer of $\mathcal{L}(\mathbf{w})$. Moreover, a *local solution* $\bar{\mathbf{w}}$ corresponds to the minimum of $\mathcal{L}(\mathbf{w})$ within an open ball centered at $\bar{\mathbf{w}}$. Finally, a *true solution* $\bar{\mathbf{w}}$ satisfies $\bar{w}_1 \odot \cdots \odot \bar{w}_N = \theta^\star$.

4 Main Results

4.1 Absence of Benign Landscape

We show that, for any arbitrary corruption probability $0 < p < 1/2$ and any number of layers $N \geq 1$, there exists at least one true solution with a strictly negative descent direction, provided that the problem is *over-parameterized*, i.e., $m \lesssim d$.

Theorem 1 (unidentifiable true solutions)**.** *Define $\mathcal{W} = \{\mathbf{w} : w_1 \odot \cdots \odot w_N = \theta^\star\}$ as the set of all true solutions of an N-layer model. For any $N \geq 1$ and $0 < p < 1/2$, the following statements hold:*

- *(Over-parameterized regime) If $m \leq 0.1d$, with probability of at least $1/16$, we have*

$$\inf_{\mathbf{w}^\star \in \mathcal{W}} \inf_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}^\star\|_\infty \leq \gamma} \{\mathcal{L}(\mathbf{w}) - \mathcal{L}(\mathbf{w}^\star)\} \lesssim -\sqrt{\frac{p_0 p}{m}} d\gamma, \tag{2}$$

for any $\gamma \lesssim t_0/\sqrt{d} \wedge 1$.

- *(Under-parameterized regime) If $m \gtrsim \frac{d}{(1-2p)^2}$, with probability of at least $1 - e^{-\Omega(d)}$, we have*

$$\inf_{\mathbf{w}^\star \in \mathcal{W}} \inf_{\mathbf{w}} \{\mathcal{L}(\mathbf{w}) - \mathcal{L}(\mathbf{w}^\star)\} \geq 0. \tag{3}$$

The above proposition unravels a sharp transition in the landscape of robust linear regression with an N-layer model: when $m \lesssim d$, some of the true solutions are likely to be non-critical points, and hence, cannot be recovered via any first-order algorithm. As soon as $m \gtrsim d$, all true solutions become global. This is in stark contrast with the recent results on the benign landscape of robust low-rank matrix recovery with ℓ_1-loss, which show that, under the so-called *restricted isometry property* (RIP), all the true solutions are global and vice versa [27, 13, 17]. The vector version of RIP is known to hold with $m = \tilde{\Omega}(k)$ samples (see e.g. [3] for a simple proof). Theorem 1 shows that, unlike the low-rank matrix recovery, RIP is *not enough* to guarantee the equivalence between the true and global solutions in deep linear models.

Theorem 1 implies that, despite their convexity, 1-layer models are *not* suitable for the robust linear regression since the set of true solutions (which is a singleton $\mathcal{W} = \{\theta^\star\}$) is unidentifiable. However, despite the existence of unidentifiable true solutions in N-layer models with $N \geq 2$, we will show that a simple SubGM converges to a *balanced* true solution, even if an arbitrarily large fraction of the measurements are corrupted with arbitrarily large noise values. This further sheds light on the desirable landscape of deeper models in the context of linear regression.

Algorithm 1 Sub-gradient Method

```
Input: Data points $\{(x_i, y_i)\}_{i=1}^m$, number of iterations $T$, the initial point $\mathbf{w}_0$, and the step-size $\{\eta^{(t)}\}_{t=0}^T$;
Output: Solution $\mathbf{w}^{(T)}$ to (1);
for $t \leq T$ do
    Select a direction $\mathbf{d}^{(t)}$ from the sub-differential $\partial\mathcal{L}(\mathbf{w}^{(t)})$ defined as:
        $\partial_{w_i}\mathcal{L}(\mathbf{w}) = \frac{1}{m}\sum_{j=1}^{m} \mathrm{Sign}\left(y_j - \left\langle \prod_k w_k, x_j \right\rangle\right) x_j \odot \prod_{k \neq i} w_k;$    (4)
    Set $\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} - \eta^{(t)}\mathbf{d}^{(t)}$;
end for
```

4.2 Convergence of Sub-gradient Method

At every iteration t, SubGM selects a direction $\mathbf{d}^{(t)}$ from the sub-differential of the ℓ_1-loss (defined as (4)), and updates the solution as $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta^{(t)}\mathbf{d}^{(t)}$; see Algorithm 1 for details. Our next two theorems characterizes the performance of SubGM with small initialization on N-layer models. We consider the cases $N = 2$ and $N \geq 3$ separately, as SubGM behaves differently on these models. We define $\kappa = \theta^\star_{\max}/\theta^\star_{\min}$ as the condition number, where $\theta^\star_{\max}$ and $\theta^\star_{\min}$ are the maximum and minimum nonzero elements of $\theta^\star$, respectively.

Theorem 2 (2-layer model)**.** *Consider the iterations of SubGM $\{\mathbf{w}^{(t)}\}_{t=0}^T$ applied to $\mathcal{L}(\mathbf{w})$ with $N = 2$ and step-size $\eta \lesssim 1$. Suppose that the initial point satisfies $w_1^{(0)} = w_2^{(0)} = \Theta(\sqrt{\alpha}\mathbf{1})$, where $0 < \alpha \lesssim d^2 m/k$. Moreover, suppose that $m \gtrsim \frac{k^2\kappa^2\log^2(m)\log(d)\log(\|\theta^\star\|/\alpha)}{(1-p)^2}$. Then, the following statements hold with probability of $1 - Ce^{-\tilde{\Omega}(k)}$:*

- ***Convergence guarantee:*** *After $\frac{1}{\eta}\log\left(\frac{1}{\alpha}\right) \lesssim \bar{T} \lesssim \frac{k^{3/2}}{\eta}\log\left(\frac{1}{\alpha}\right)$ iterations, we have*

$$\left\|w_1^{(\bar{T})} \odot w_2^{(\bar{T})} - \theta^\star\right\| \lesssim \eta\theta^\star_{\max} \vee \sqrt{d^2 m\alpha}^{\,1-\tilde{\Theta}\left(\frac{k^2}{\sqrt{(1-p)^2 m}}\right)}.$$

- ***Balanced property:*** *For every $0 \leq t \leq \bar{T}$, we have*

$$\left\|w_1^{(t)} - w_2^{(t)}\right\|_\infty \lesssim \alpha^{0.5-\tilde{\Theta}\left(\frac{k^2}{\sqrt{(1-p)^2 m}}\right)}.$$

- ***Long escape time:*** *For every $\bar{T} \leq t \leq \sqrt{\frac{m(1-p)^2}{k}}\bar{T}$, we have*

$$\left\|w_1^{(t)} \odot w_2^{(t)} - \theta^\star\right\| \lesssim \eta\theta^\star_{\max} \vee \sqrt{d^2 m\alpha}^{\,0.5-\tilde{\Theta}\left(\frac{k^2}{\sqrt{(1-p)^2 m}}\right)}.$$

Furthermore, if $m \gtrsim d\log(m)/(1-p)^2$, with probability of $1 - Ce^{-\tilde{\Omega}(k)}$ and for every $t \geq \bar{T}$, we have

$$\left\|w_1^{(t)} \odot w_2^{(t)} - \theta^\star\right\| \lesssim \eta\theta^\star_{\max} \vee \sqrt{d^2 m\alpha}^{\,1-\tilde{\Theta}\left(\frac{k^2}{\sqrt{m(1-p)^2}}\right)}\left(1 - \Omega\left(\eta/\sqrt{d}\right)\right)^{t-\bar{T}}.$$

We provide the main idea behind the proof of Theorem 2 in Section 5. The formal proof can be found in the appendix. A few observations are in order based on Theorem 2. First, for any $\epsilon > 0$, SubGM is guaranteed to satisfy $\left\| w_1^{(t)} \odot w_2^{(t)} - \theta^\star \right\| \lesssim \epsilon$ after $\mathcal{O}((1/\epsilon)\log(d/\epsilon))$ iterations, provided that $\eta = \Theta(\epsilon)$ and $\alpha = \epsilon^2/(d^2m)$. Based on our numerical results (provided in the appendix), we believe that it is possible to establish a linear convergence for SubGM with a geometric step-size; a rigorous verification of this conjecture is considered as future work. Second, although SubGM converges to a vicinity of a true solution quickly, it will stay there for a significantly longer time—in particular, $\sqrt{m(1-p)^2/k}$ times longer than its initial convergence time. Such behavior is also exemplified in our simulations (see Figure 1e). After this escape time, the algorithm may slowly converge to an *overfitted* solution with a better training loss. Moreover, if $m \gtrsim d$, SubGM will continuously converge to a true solution at an exponential rate, and it will never diverge. Finally, Theorem 2 shows that SubGM implicitly favors balanced solutions, i.e. solutions whose factors have similar magnitudes. Combined with the convergence result of SubGM, we immediately conclude that SubGM converges to a particular solution of the form $(\sqrt{\theta^\star}, \sqrt{\theta^\star})$. Therefore, the solution found by SubGM will enjoy the same (approximate) sparsity pattern as $\theta^\star$.

Theorem 3 (N-layer models)**.** *Consider the iterations of SubGM $\{\mathbf{w}^{(t)}\}_{t=0}^T$ applied to $\mathcal{L}(\mathbf{w})$ with $N \geq 3$ and step-size $\eta \lesssim N^{-1}\kappa^{-\frac{N-2}{N}}$. Suppose that the initial point satisfies $w_j^{(0)} = \Theta(\alpha^{1/N}\mathbf{1})$, where $0 < \alpha \lesssim d^2m/k$. Moreover, suppose that $m \gtrsim \frac{k^2\kappa^4\log^2(m)\log(d)\log(\|\theta^\star\|/\alpha)}{(1-p)^2}$. Then, the following statements hold with probability of $1 - Ce^{-\tilde{\Omega}(k)}$:*

- ***Convergence guarantee:*** *After $\frac{1}{N\eta}\alpha^{-\frac{N-2}{N}} \lesssim \bar{T} \lesssim \frac{k^{3/2}}{N\eta}\alpha^{-\frac{N-2}{N}}$ iterations, we have*

$$\left\| \prod w_i^{(\bar{T})} - \theta^\star \right\| \lesssim N\eta\theta^\star_{\max} \vee \sqrt{d^2m\alpha}.$$

- ***Balanced property:*** *For every $0 \leq t \leq \bar{T}$, we have*

$$\left| w_{i,l}^{(t)} - w_{j,l}^{(t)} \right| = \mathcal{O}\left(\alpha^{1/N}\right), \qquad \text{for } 1 \leq i < j \leq N, l : \theta_l^\star = 0,$$
$$\left| w_{i,l}^{(t)} - w_{j,l}^{(t)} \right| = \tilde{\mathcal{O}}\left(\sqrt[N]{\theta_l^\star}\sqrt{k^3/m}\right), \quad \text{for } 1 \leq i < j \leq N, l : \theta_l^\star \neq 0.$$

- ***Long escape time:*** *For every $\bar{T} \leq t \leq \sqrt{\frac{m(1-p)^2}{k}}\bar{T}$, we have*

$$\left\| \prod w_i^{(t)} - \theta^\star \right\| \lesssim N\eta\theta^\star_{\max} \vee \sqrt{d^2m\alpha},$$

Furthermore, if $m \gtrsim d^{\frac{2N-2}{N}}\log(m)/(1-p)^2$, with probability of at least $1 - Ce^{-\tilde{\Omega}(k)}$ and for every $t > \bar{T}$, we have

$$\left\| \prod w_i^{(t)} - \theta^\star \right\| \lesssim N\eta\theta^\star_{\max} \vee \left(\frac{\sqrt{d^2m}\alpha}{\sqrt{d^2m}\alpha N\eta d^{-\frac{N-1}{N}}(t-\bar{T})+1} \right)^{\frac{N}{N-2}}.$$

The proof of this theorem can be found in the appendix. Theorem 3 sheds light on an important benefit of N-layer models with $N \geq 3$ compared to 2-layer models: for sufficiently small step-size, deeper models improve the generalization error by a factor of $(1/\alpha)^{\tilde{\Theta}\left(k^2/\sqrt{((1-p)^2m)}\right)}$. This improvement is particularly significant when both α and m are small. However, such improvement comes at the expense of a slower convergence rate. In particular, after setting $\eta = \Theta(\epsilon/N)$, and $\alpha = \epsilon/\sqrt{d^2 m}$, SubGM needs $\mathcal{O}\left((1/\epsilon)^{1+\frac{N-2}{N}}\right)$ iterations to obtain an ϵ-accurate solution. Evidently, the convergence rate deteriorates with N, ultimately approaching $\mathcal{O}\left(1/\epsilon^2\right)$ for infinitely deep models. This can be observed in practice: Figures 1e and 1f show that 3-layer model enjoys a better generalization error compared to 2-layer model, but suffers from a slower convergence rate. This slower convergence rate also manifests itself in a more stable behavior of the algorithm: for deeper models, SubGM stays close to the ground truth for a longer time. Finally, the balanced property of the solution obtained via SubGM extends to N-layer models. In particular, SubGM converges to a particular solution of the form $(\sqrt[N]{\theta^\star}, \dots, \sqrt[N]{\theta^\star})$, thereby inheriting the same sparsity pattern as $\theta^\star$.

4.3 Local Landscape Around Balanced Solution

In the previous section, we showed that SubGM converges to a balanced solution. In this section, we characterize the local landscape around this balanced solution, proving that it becomes flatter for deeper models.

Theorem 4 (flatness around balanced solution)**.** *Suppose that* $k\log(d)/(1-2p)^2 \lesssim m \leq 0.1d$ *and* $p < 1/2$*. Let* $\mathbf{w}^\star = (\sqrt[N]{\theta^\star}, \dots, \sqrt[N]{\theta^\star})$*. Then, for any* $N \geq 2$ *and* $\gamma \leq t_0/\sqrt{d} \wedge 1$*, the following statements hold:*

- *With probability at least* $1 - e^{-\Omega(k)}$*, we have*

$$\inf_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}^\star\|_\infty \leq \gamma} \left\{\mathcal{L}(\mathbf{w}^\star) - \mathcal{L}(\mathbf{w})\right\} \gtrsim -\frac{d}{\sqrt{m}}\gamma^N.$$

- *With probability at least* $1/16$*, we have*

$$\inf_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}^\star\|_\infty \leq \gamma} \left\{\mathcal{L}(\mathbf{w}^\star) - \mathcal{L}(\mathbf{w})\right\} \lesssim -\sqrt{p_0 p}\frac{d}{\sqrt{m}}\gamma^N.$$

Theorem 4 shows that, within a γ-neighborhood of $\mathbf{w}^\star$, the most descent direction from $\mathbf{w}^\star$ can reduce the loss by at most $\mathcal{O}\left(d/\sqrt{m}\cdot\gamma^N\right)$, which decreases exponentially with N. Moreover, in the noisy setting, the above theorem implies that $\mathbf{w}^\star$ is likely to be neither local nor global minimum, since it has a descent direction. However, the flatness of the landscape around $\mathbf{w}^\star$ enables SubGM to stay close to the balanced solution for a long time.

Remark 1. *Note that the choice of* ℓ_∞*-ball for the perturbation set is to ensure that the size of the possible perturbations per layer remains independent of the depth of the model. This is indeed crucial to ensure a fair comparison between models with different depths: alternative choices of the perturbation set, such as* ℓ_q*-ball with* $1 \leq q < \infty$ *(e.g.* ℓ_2*-ball) would shrink the size of the feasible per-layer perturbations with* N*, thereby leading to an unfair advantage to deeper models.*

5 Proof Techniques

At the crux of our proof technique for Theorems 2 and 3 lies the following decomposition of the sub-differential:

$$\partial\mathcal{L}(\mathbf{w}) = \underbrace{\xi\cdot\partial\bar{\mathcal{L}}(\mathbf{w})}_{\text{expected subdiff.}} + \underbrace{\left(\partial\mathcal{L}(\mathbf{w}) - \xi\cdot\partial\bar{\mathcal{L}}(\mathbf{w})\right)}_{\text{subdiff. deviation}}, \quad \text{for some strictly positive } \xi.$$

In the above decomposition, $\bar{\mathcal{L}}(\mathbf{w})$ is called *expected loss*, and is defined as $\bar{\mathcal{L}}(\mathbf{w}) = \|w_1 \odot \cdots \odot w_N - \theta^\star\|$. As will be shown later, $\bar{\mathcal{L}}(\mathbf{w})$ captures the expected behavior of the empirical loss $\mathcal{L}(\mathbf{w})$. To analyze the behavior of SubGM on $\mathcal{L}(\mathbf{w})$, we first consider the ideal scenario, where $\mathcal{L}(\mathbf{w})$ coincides with its expectation. Then, we extend our analysis to the general case by controlling the sub-differential deviation. In particular, we show that the desirable convergence properties of SubGM extends to $\mathcal{L}(\mathbf{w})$, provided that its sub-differentials are "direction-preserving", i.e., $\mathbf{d} \approx \xi\bar{\mathbf{d}}$, for every $\mathbf{d} \in \partial\mathcal{L}(\mathbf{w}), \bar{\mathbf{d}} \in \partial\bar{\mathcal{L}}(\mathbf{w})$ and some $\xi > 0$. To formalize this idea, we first provide a more concise characterization of $\partial\mathcal{L}(\mathbf{w})$:

$$\partial_{w_i}\mathcal{L}(\mathbf{w}) = \left\{ q \odot \prod_{k\neq i} w_k : q \in \mathcal{Q}\left(\theta^\star - \prod_k w_k\right)\right\}, \text{ where } \mathcal{Q}(z) = \frac{1}{m}\sum_{i=1}^m \operatorname{Sign}\left(\langle x_i, z\rangle + \epsilon_i\right) x_i.$$

Definition 1 (approximately sparse vectors)**.** *We say a vector $v \in \mathbf{R}^d$ is (k,ϑ)-approximately sparse if there exists a vector u, such that $\|u\|_0 \le k$ and $\|u - v\| \le \vartheta$.*

Proposition 1 (direction-preserving property)**.** *Suppose that $m \gtrsim \frac{k\log^2(m)\log(d)\log(R/\vartheta)}{(1-p)^2\delta^2}$ for some $R, \vartheta, \delta > 0$. Then, with probability of at least $1 - Ce^{-\Omega(m\delta^2)}$, the following inequality holds for any $q \in \mathcal{Q}(z)$ and any (k,ϑ)-approximately sparse vector z that satisfies $\sqrt{dm/k}\vartheta\log(1/\vartheta) \lesssim \|z\| \le R$:*

$$\left\| q - \sqrt{\frac{2}{\pi}}\left(1 - p + p\mathbb{E}\left[e^{-\epsilon^2/(2\|z\|)}\right]\right)\frac{z}{\|z\|}\right\|_\infty \le \delta. \tag{5}$$

Moreover, if $m \gtrsim \frac{d\log(m)}{(1-p)^2\delta^2}$, with probability of $1 - Ce^{-\Omega(m\delta^2)}$, (5) holds for every $z \in \mathbb{R}^d$.

Proposition 1 is analogous to *Sign-RIP* condition introduced in [30, 29] for the robust low-rank matrix recovery, and is at the heart of our proofs for Theorems 2 and 3. Suppose that $\theta^\star - \prod_k w_k$ is a (k,ϑ)-approximately sparse and satisfies (5). Then, we have $\left\|\mathbf{d} - \bar{\mathbf{d}}\right\|_\infty \le \left(\max_i\left\{\prod_{k\neq i} w_k\right\}\right)\delta$, which in turn provides an upper bound on the sub-differential deviation.

5.1 Proof Sketch of Theorem 2

To streamline the presentation, here we only provide simplified versions of our key ideas, which inevitably lead to looser guarantees. To streamline the proof, we assume that $\theta^\star_1 \ge \cdots \ge \theta^\star_k > \theta^\star_{k+1} = \cdots = \theta^\star_d = 0$. Moreover, for simplicity of notation, we denote $u = w_1$ and $v = w_2$. Consider the following decomposition:

$$u \odot v = [\underbrace{u_1v_1 \ldots u_kv_k}_{S}\ \underbrace{u_{k+1}v_{k+1} \ldots u_dv_d}_{E}]^\top. \tag{6}$$

The vectors S and E are called *signal* and *residual terms*, respectively. Evidently, we have $u \odot v = \theta^\star$ if and only if $S = [\theta^\star_1, \ldots, \theta^\star_k]^\top$ and $E = 0$. Based on this observation, our goal is to show that the signal term converges to $[\theta^\star_1, \ldots, \theta^\star_k]^\top$ exponentially fast, while the error term remains small throughout the solution trajectory.

Lemma 1 (signal dynamic; informal)**.** *Suppose that* (5) *holds for* $z = \theta^\star - u^{(t)} \odot v^{(t)}$*, and* $\left\|\theta^\star - u^{(t)} \odot v^{(t)}\right\| \gtrsim \eta \left\|\theta^\star\right\|$*. Then, we have*

$$u_i^{(t+1)} v_i^{(t+1)} \geq \left(1+2\eta\left(\frac{\theta_i^\star - u_i^{(t)} v_i^{(t)}}{\left\|u^{(t)} \odot v^{(t)} - \theta^\star\right\|} + \delta_i\right)\right) u_i^{(t)} v_i^{(t)}, \tag{7}$$

for some $|\delta_i| \leq \delta$ *and every* $i = 1, \ldots, k$*.*

Lemma 2 (residual dynamic; informal)**.** *Suppose that* (5) *holds for* $z = \theta^\star - u^{(t)} \odot v^{(t)}$*, and* $\left\|\theta^\star - u^{(t)} \odot v^{(t)}\right\| \gtrsim \eta \left\|\theta^\star\right\|$*. Then, we have*

$$\left(u_i^{(t+1)}\right)^2 + \left(v_i^{(t+1)}\right)^2 \leq (1+\mathcal{O}(\eta\delta))\left(\left(u_i^{(t)}\right)^2 + \left(v_i^{(t)}\right)^2\right), \tag{8}$$

for every $i = k+1, \ldots, d$*.*

Lemma 3 (difference dynamic; informal)**.** *Suppose that* (5) *holds for* $z = \theta^\star - u^{(t)} \odot v^{(t)}$*, and* $\left\|\theta^\star - u^{(t)} \odot v^{(t)}\right\| \gtrsim \eta \left\|\theta^\star\right\|$*. Then, we have*

$$u_i^{(t+1)} - v_i^{(t+1)} = \left(u_i^{(t)} - v_i^{(t)}\right)\left(1 - \eta\frac{\theta_i^\star - u_i^{(t)} v_i^{(t)}}{\left\|u^{(t)} \odot v^{(t)} - \theta^\star\right\|} + \eta\delta_i\right), \tag{9}$$

for some $|\delta_i| \leq \delta$ *and every* $i = 1, \ldots, d$*.*

Convergence guarantee. For any fixed $i = 1, \ldots, k$, we show that $u_i^{(t)} v_i^{(t)} = \theta_i^\star \pm \mathcal{O}(\delta)\left\|\theta^\star\right\|$ after $\mathcal{O}(\|\theta^\star\|/(\eta\theta_i^\star)\log(1/\alpha))$ iterations. To see this, suppose that T_i is the largest iteration such that $u_i^{(t)} v_i^{(t)} \leq \theta_i^\star$ for every $t \leq T_i$. Moreover, suppose that $\left\|u^{(t)} \odot v^{(t)}\right\| \leq C\left\|\theta^\star\right\|$, for sufficiently large C (this is proven in the appendix). Under these assumptions, (7) reduces to

$$u_i^{(t+1)} v_i^{(t+1)} \geq \left(1 + \Omega(1)\frac{\eta\theta_i^\star}{\|\theta^\star\|}\right) u_i^{(t)} v_i^{(t)}. \tag{10}$$

which implies that $T_i \lesssim \|\theta^\star\|/(\eta\theta_i^\star)\log(1/\alpha)$. For any $t > T_i$, define $y_i^{(t)} = \theta_i^\star - u_i^{(t)} v_i^{(t)}$. One can write

$$y_i^{(t+1)} \leq \left(1 - \Omega(1)\frac{\eta\theta_i^\star}{\|\theta^\star\|}\right) y_i^{(t)} + \eta\delta\theta_i^\star. \tag{11}$$

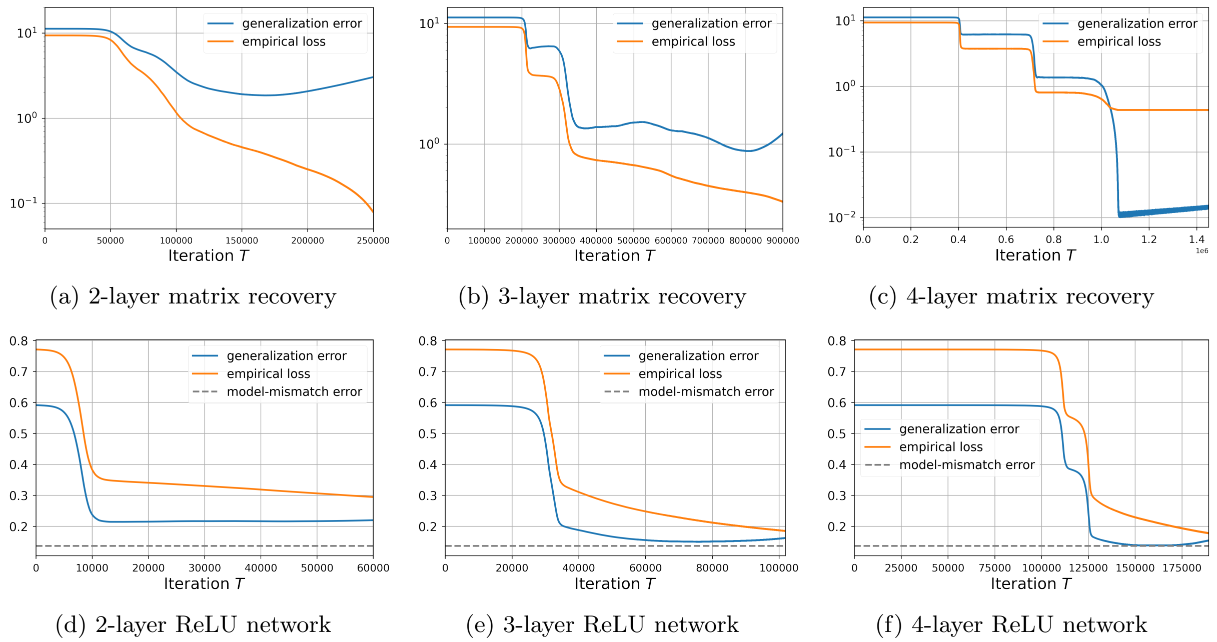
Hence, with additional $\mathcal{O}\left(\left\|\theta^\star\right\|/(\eta\theta_1^\star)\right)$ iterations, we have $u_i^{(t)} v_i^{(t)} = \theta_i^\star \pm \mathcal{O}(\delta)\left\|\theta^\star\right\|$. On the other hand, Lemma 2 implies that, for any $i = k+1, \ldots, d$ and $t \lesssim \|\theta^\star\|/(\eta\theta_k^\star)\log(1/\alpha)$, we have

$$\left(u_i^{(t)}\right)^2 + \left(v_i^{(t)}\right)^2 \lesssim \alpha\left(1+\mathcal{O}(\eta\delta)\right)^{\mathcal{O}\left(\frac{\|\theta^\star\|}{\eta\theta_k^\star}\log\left(\frac{1}{\alpha}\right)\right)} \lesssim \alpha^{1-\mathcal{O}\left(\sqrt{k}\kappa\delta\right)},$$

where $\kappa = \theta_1^\star/\theta_k^\star$ is the condition number of $\theta^\star$. Combining the above dynamics, we have

$$\left\|u^{(t)} \odot v^{(t)} - \theta^\star\right\| \lesssim \eta\left\|\theta^\star\right\| \vee \sqrt{k}\left\|\theta^\star\right\|\delta \vee \sqrt{d}\alpha^{1-\mathcal{O}\left(\sqrt{k}\kappa\delta\right)}.$$

In the appendix, we provide a more refined analysis that relaxes the dependency of the final error on δ and κ.



(a) 2-layer matrix recovery

(b) 3-layer matrix recovery

(c) 4-layer matrix recovery

(d) 2-layer ReLU network

(e) 3-layer ReLU network

(f) 4-layer ReLU network

Figure 2: **Deep matrix recovery (first row).** The ground truth $X^\star \in \mathbf{R}^{20\times 20}$ with $\operatorname{rank}(X^\star) = 3$ is chosen randomly. The elements of the measurement matrices are selected from $\mathcal{N}(0,1)$, and the sample size is set to $m = 180$. The corruption probability is set to $p = 0.05$ with distribution $\mathcal{N}(0,100)$. We use SubGM with step-size $\eta = 0.001$ and Gaussian initialization with an initialization scale $\alpha = 1\times 10^{-3}$. **ReLU models (second row).** The samples are chosen from $y_i = \sin(\theta^{\star\top} x_i) + \epsilon_i$ where $\theta^\star \in \mathbf{R}^{50}$ is randomly generated with $\|\theta^\star\|_0 = 2$, $x_i \sim \mathcal{N}(0, I_{50})$, and $\epsilon_i \sim \mathcal{N}(0,25)$ with corruption probability $p = 0.05$. The sample size is set to be $m = 1500$. We use SubGM with step-size $\eta = 0.001$ and Gaussian initialization $\mathcal{N}(0, \alpha^{2/N}/d)$.

Long escape time. We show in the appendix that after the first stage, the residual becomes the dominant term in the final error. This together with Lemma 2 implies that, for every $t \lesssim \frac{\|\theta^\star\|}{\eta\theta_k^\star\sqrt{\delta}}\log(1/\alpha)$, we have $\|E\| \lesssim \sqrt{d}\alpha^{1-\sqrt{k}\kappa\sqrt{\delta}}$.

Balanced property. We have $u_i^{(t)}v_i^{(t)} \le \theta_i^\star$ for every $i \in [k]$, and $|u_i^{(t)}v_i^{(t)}| \lesssim \alpha^{1-\mathcal{O}(\sqrt{k}\kappa\delta)}$ for every $i = k+1,\dots,d$. Therefore, Lemma 3 can be invoked to verify $\left|u_i^{(t+1)} - v_i^{(t+1)}\right| \le (1+\mathcal{O}(\eta\delta))\left|u_i^{(t)} - v_i^{(t)}\right|$. This in turn leads to

$$\left|u_i^{(t)} - v_i^{(t)}\right| \lesssim \sqrt{\alpha}\,(1+\mathcal{O}(\eta\delta))^{\mathcal{O}\left(\frac{\|\theta^\star\|}{\eta\theta_k^\star}\log\left(\frac{1}{\alpha}\right)\right)} \lesssim \alpha^{0.5-\mathcal{O}(\sqrt{k}\kappa\delta)}.$$

6 Numerical Experiments: Beyond Linear Regression

In this section, we empirically verify that the benefits of depth extend to the robust variants of deep matrix recovery and ReLU networks with ℓ_1-loss.

Deep Matrix Recovery. In low-rank matrix recovery, the goal is to recover a low-rank matrix $X^\star \in \mathbf{R}^{d\times d}$, from a limited number of noisy measurements of the form $y_i = \langle A_i, X^\star \rangle + \epsilon_i$. To recover $X^\star$, we consider a deep factorized model of the form $W_1 W_2 \dots W_N$, where $W_i \in \mathbf{R}^{d\times d}$ for $i = 1, \dots, N$, and minimize the ℓ_1-loss $(1/m)\sum_{i=1}^m |y_i - \langle A_i, W_1 W_2 \cdots W_N \rangle|$ via SubGM. When $N = 2$, the above model reduces to the famous Burer-Monteiro approach [9]. We assume that 5% of the measurements are grossly corrupted with noise. The first row of Figure 5 shows the performance of SubGM on 2-, 3-, and 4-layer models. It can be seen that the 4-layer model outperforms shallower models, achieving a generalization error that is proportional to the step-size.

Deep ReLU Network on Synthetic Dataset. As another experiment, we analyze the effect of depth on the performance of SubGM with ReLU networks and ℓ_1-loss. Given an input $x \in \mathbf{R}^d$, the output of an N-layer ReLU network is defined as $f_{\mathbf{W}}(x) = W_N \sigma\left(W_{N-1} \cdots \sigma\left(W_1 x\right) \cdots\right)$, where $W_1 \in \mathbf{R}^{m\times d}, W_2, \cdots, W_{N-1} \in \mathbf{R}^{m\times m}$, and $W_N \in \mathbb{R}^{1\times m}$. Moreover, $\sigma(x) = \max\{0, x\}$ is the ReLU activation function. Given the true function $f^\star(x) = \sin(\theta^{\star\top} x)$, our goal is to train a ReLU model to approximate $f^\star$ as accurately as possible. To this goal, we minimize the ℓ_1-loss $(1/m)\sum_{i=1}^m |y_i - f_{\mathbf{W}}(x_i)|$. The second row of Figure 5 illustrates the performance of SubGM. It is worth noting that, unlike robust linear regression and deep matrix recovery, there always exists a non-diminishing model-mismatch error between the true and considered ReLU model (shown as a dashed line). Nonetheless, SubGM can achieve this model-mismatch error on a 4-layer ReLU model with only 1500 samples, even if 5% of the measurements are corrupted with large noise.

Deep ReLU Network on CIFAR Dataset We verify that the desirable performance of SubGM with ℓ_1-loss can be extended to its stochastic variant with mini-batches on CIFAR-10 and CIFAR-100 [25], outperforming cross-entropy (CE) loss, which is considered as one of the most suitable loss functions for CIFAR datasets. To show this, we use standard ResNet architectures [22] with ℓ_1-loss and compare it with the cross-entropy loss on noisy CIFAR datasets, where we randomize the labels of 10% of the training dataset. For CIFAR-100 experiment, we use the "loss scaling" trick introduced in [23]. The training details are deferred to Section A.3. The best test accuracy for both CIFAR-10 and CIFAR-100 is reported in Table 1. One can see that ℓ_1-loss outperforms cross-entropy loss significantly, demonstrating that our framework may be extended to more realistic settings. Moreover, we do observe that the deeper model performs better on CIFAR-100, which aligns with our theoretical result. Based on our simulations, an interesting and important future direction would be to study the optimization landscape of ℓ_1-loss with more general neural network architectures.

Method	CIFAR-10			CIFAR-100		
	ResNet-18	ResNet-34	ResNet-50	ResNet-18	ResNet-34	ResNet-50
CE loss	91.52%	91.53%	90.87%	70.17%	71.22%	71.30%
ℓ_1-loss	**94.16%**	**93.13%**	**92.68%**	**73.69%**	**74.27%**	**75.19%**

Table 1: Test accuracy for ResNet-18, 34, 50 on CIFAR-10 and CIFAR-100 datasets with noise.

7 Conclusion

Modern problems in machine learning are naturally nonconvex but can be solved reasonably well in practice. To explain this, a recent body of work has postulated that many optimization problems in machine learning are "convex-like", i.e., they are devoid of spurious local minima. Our work shows that such global property is too restrictive to hold even in the context of linear regression, and instead propose a more refined *trajectory analysis* to better capture the landscape of the problem around the solution trajectory. We show that convex models may be fundamentally ill-suited for linear models, and deeper models–despite their nonconvexity–have provably better optimization landscape around the solution trajectory. Empirically, we show that our analysis may extend beyond linear regression; a formal verification of this conjecture is considered as an enticing challenge for future research.

Acknowledgements

We thank Richard Y. Zhang, Cédric Jósz, and Tiffany Wu for helpful discussions and feedback. We are also thankful for an anonymous reviewer for pointing out the relationship between the perturbation ball and the depth of linear models. This research is supported, in part, by NSF Award DMS-2152776, ONR Award N00014-22-1-2127, MICDE Catalyst Grant, MIDAS PODS grant and Startup funding from the University of Michigan.

References

[1] Zeyuan Allen-Zhu and Yuanzhi Li. Backward feature correction: How deep learning performs deep learning. arXiv preprint arXiv:2001.04413, 2020.

[2] Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. Advances in Neural Information Processing Systems, 32:7413–7424, 2019.

[3] Richard Baraniuk, Mark Davenport, Ronald DeVore, and Michael Wakin. A simple proof of the restricted isometry property for random matrices. Constructive Approximation, 28(3):253–263, 2008.

[4] Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. Proceedings of the National Academy of Sciences, 117(48):30063–30070, 2020.

[5] Dimitris Bertsimas, Angela King, and Rahul Mazumder. Best subset selection via a modern optimization lens. The annals of statistics, 44(2):813–852, 2016.

[6] Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Global optimality of local search for low rank matrix recovery. arXiv preprint arXiv:1605.07221, 2016.

[7] Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. The Annals of statistics, 37(4):1705–1732, 2009.

[8] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258, 2021.

[9] Samuel Burer and Renato DC Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. Mathematical Programming, 95(2):329–357, 2003.

[10] Emmanuel Candes and Terence Tao. The dantzig selector: Statistical estimation when p is much larger than n. The annals of Statistics, 35(6):2313–2351, 2007.

[11] Hung-Hsu Chou, Johannes Maly, and Holger Rauhut. More is less: Inducing sparsity via overparameterization. arXiv preprint arXiv:2112.11027, 2021.

[12] Damek Davis and Dmitriy Drusvyatskiy. Stochastic subgradient method converges at the rate $o(k^{-1/4})$ on weakly convex functions. arXiv preprint arXiv:1802.02988, 2018.

[13] Lijun Ding, Liwei Jiang, Yudong Chen, Qing Qu, and Zhihui Zhu. Rank overspecified robust matrix recovery: Subgradient method and exact recovery. Advances in Neural Information Processing Systems, 34, 2021.

[14] Simon Du and Wei Hu. Width provably matters in optimization for deep linear neural networks. In International Conference on Machine Learning, pages 1655–1664. PMLR, 2019.

[15] Simon S Du, Wei Hu, and Jason D Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. arXiv preprint arXiv:1806.00900, 2018.

[16] Ronen Eldan and Ohad Shamir. The power of depth for feedforward neural networks. In Conference on learning theory, pages 907–940. PMLR, 2016.

[17] Salar Fattahi and Somayeh Sojoudi. Exact guarantees on the absence of spurious local minima for non-negative rank-1 robust principal component analysis. Journal of machine learning research, 2020.

[18] Jerome Friedman, Trevor Hastie, Robert Tibshirani, et al. The elements of statistical learning, volume 1. Springer series in statistics New York, 2001.

[19] Daniel Gissin, Shai Shalev-Shwartz, and Amit Daniely. The implicit bias of depth: How incremental learning drives generalization. arXiv preprint arXiv:1909.12051, 2019.

[20] Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Implicit bias of gradient descent on linear convolutional networks. arXiv preprint arXiv:1806.00468, 2018.

[21] Jeff Z HaoChen, Colin Wei, Jason Lee, and Tengyu Ma. Shape matters: Understanding the implicit bias of the noise covariance. In Conference on Learning Theory, pages 2315–2357. PMLR, 2021.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[23] Like Hui and Mikhail Belkin. Evaluation of neural architectures trained with square loss vs cross-entropy in classification tasks. In International Conference on Learning Representations, 2020.

[24] Kenji Kawaguchi. Deep learning without poor local minima. arXiv preprint arXiv:1605.07110, 2016.

[25] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[26] Jiangyuan Li, Thanh Nguyen, Chinmay Hegde, and Ka Wai Wong. Implicit sparse regularization: The impact of depth and early stopping. Advances in Neural Information Processing Systems, 34, 2021.

[27] Xiao Li, Zhihui Zhu, Anthony Man-Cho So, and Rene Vidal. Nonconvex robust low-rank matrix recovery. SIAM Journal on Optimization, 30(1):660–686, 2020.

[28] Zhiyuan Li, Yuping Luo, and Kaifeng Lyu. Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning. arXiv preprint arXiv:2012.09839, 2020.

[29] Jianhao Ma and Salar Fattahi. Sign-rip: A robust restricted isometry property for low-rank matrix recovery. arXiv preprint arXiv:2102.02969, 2021.

[30] Jianhao Ma and Salar Fattahi. Global convergence of sub-gradient method for robust matrix recovery: Small initialization, noisy measurements, and over-parameterization. arXiv preprint arXiv:2202.08788, 2022.

[31] Rahul Mazumder, Peter Radchenko, and Antoine Dedieu. Subset selection with shrinkage: Sparse linear modeling when the snr is low. arXiv preprint arXiv:1708.03288, 2017.

[32] Roman Novak, Yasaman Bahri, Daniel A Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Sensitivity and generalization in neural networks: an empirical study. arXiv preprint arXiv:1802.08760, 2018.

[33] Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. IEEE transactions on information theory, 57(10):6976–6994, 2011.

[34] Itay Safran and Jason D Lee. Optimization-based separations for neural networks. arXiv preprint arXiv:2112.02393, 2021.

[35] Matus Telgarsky. Representation benefits of deep feedforward networks. arXiv preprint arXiv:1509.08101, 2015.

[36] Matus Telgarsky. Benefits of depth in neural networks. In Conference on learning theory, pages 1517–1539. PMLR, 2016.

[37] Sara Van de Geer. The deterministic lasso. Seminar für Statistik, Eidgenössische Technische Hochschule (ETH) Zürich, 2007.

[38] Aad W Van Der Vaart, Adrianus Willem van der Vaart, Aad van der Vaart, and Jon Wellner. Weak convergence and empirical processes: with applications to statistics. Springer Science & Business Media, 1996.

[39] Tomas Vaskevicius, Varun Kanade, and Patrick Rebeschini. Implicit regularization for optimal sparse recovery. Advances in Neural Information Processing Systems, 32:2972–2983, 2019.

[40] Frederi G Viens and Andrew B Vizcarra. Supremum concentration inequality and modulus of continuity for sub-nth chaos processes. Journal of Functional Analysis, 248(1):1–26, 2007.

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

[42] Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In Conference on Learning Theory, pages 3635–3673. PMLR, 2020.

[43] Peng Zhao, Yun Yang, and Qiao-Chu He. Implicit regularization via hadamard product over-parametrization in high-dimensional linear regression. arXiv preprint arXiv:1903.09367, 2019.

A Additional Experiments

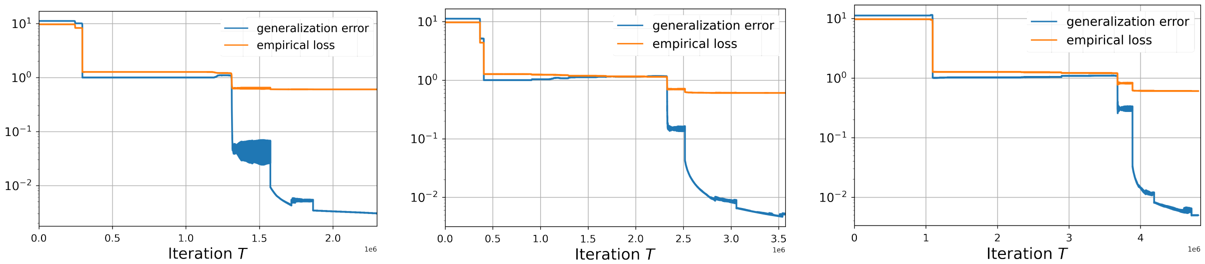
In this section, we provide additional experiments on the performance of SubGM on deep models. Our goal is to verify our theoretical results and show the benefits of both small initialization and geometric step-size. Moreover, we show that the desirable performance of SubGM can be observed in its stochastic variant, as well as for different architectures of ResNets with ℓ_1-loss, and more realistic CIFAR-10 dataset. All simulations are run on a desktop computer with an Intel Core i9 3.50 GHz CPU and 128GB RAM. The reported results are for an implementation in Python.

A.1 Deeper Linear Models

In this experiment, we study the performance of SubGM on deeper models ($N = 4, 5, 6$). To accelerate the training process, we first use a large step-size $\eta_1 = 1 \times 10^{-3}$, and then progressively apply smaller step-sizes $\eta_2 = 1 \times 10^{-4}$ and $\eta_3 = 1 \times 10^{-5}$ as the training loss continues to decay. As shown in Figure 3, deeper models share similar generalization error, outperforming 1- and 2-layer models presented in Figure 1.

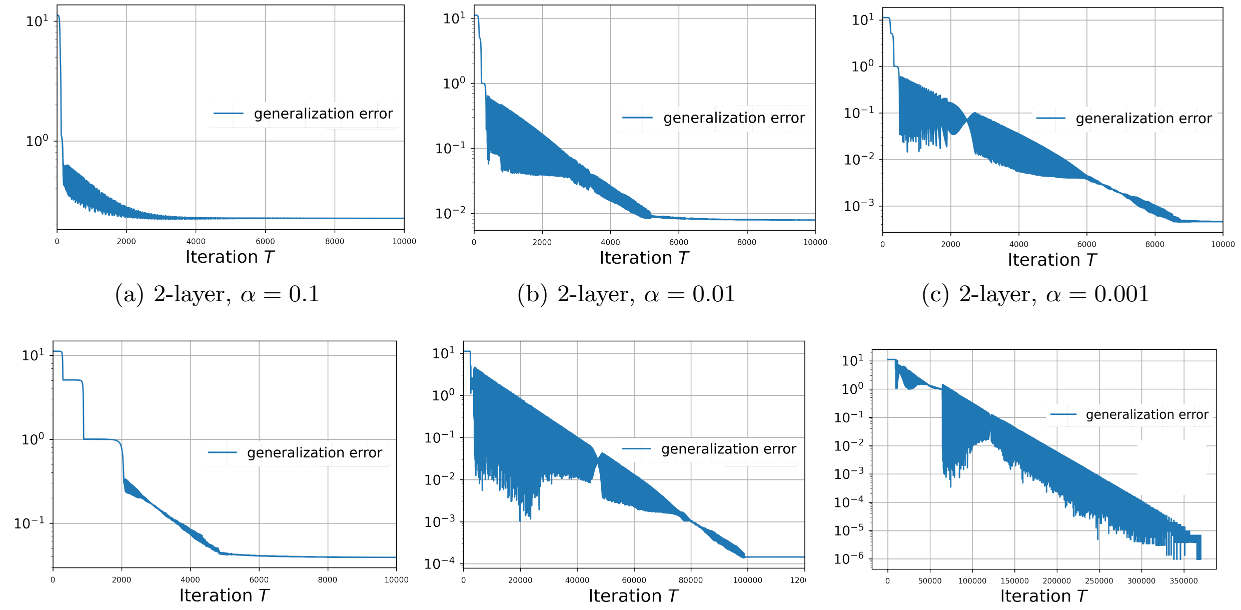
A.2 Geometric Step-size

As shown in our simulations, training N-layer models may require millions of iterations even on a small synthetic dataset. As proven in Theorem 3, the training process may become even slower for deeper models. In this experiment, we explore the performance of geometric (i.e., exponentially decaying) step-size on the same dataset. SubGM with a geometric step-size has been widely used for the optimization of ℓ_1-loss [27, 29], and more general sharp weakly convex functions [12]. Figure 4 shows that a geometric steps-size can lead to a 1000-fold reduction in the required number of iterations. Moreover, a geometric step-size improves the convergence rate to linear. The theoretical justification of this improvement is left as an enticing challenge for future research. Finally, it can be observed that SubGM with geometric step-size performs surprisingly well on deeper models, achieving a generalization errors in the order of 10^{-6}. This further supports the benefits of the depth.



(a) 4-layer (b) 5-layer (c) 6-layer

Figure 3: The optimization trajectories of deep models ($N = 4, 5, 6$).



(a) 2-layer, $\alpha = 0.1$

(b) 2-layer, $\alpha = 0.01$

(c) 2-layer, $\alpha = 0.001$

(d) 3-layer, $\alpha = 0.1$

(e) 3-layer, $\alpha = 0.01$

(f) 3-layer, $\alpha = 0.001$

Figure 4: The optimization trajectories of 2 and 3-layer models with different initialization size $\alpha = 0.1, 0.01, 0.001$ using exponentially decayed step-size.

A.3 Experiments on CIFAR Dataset

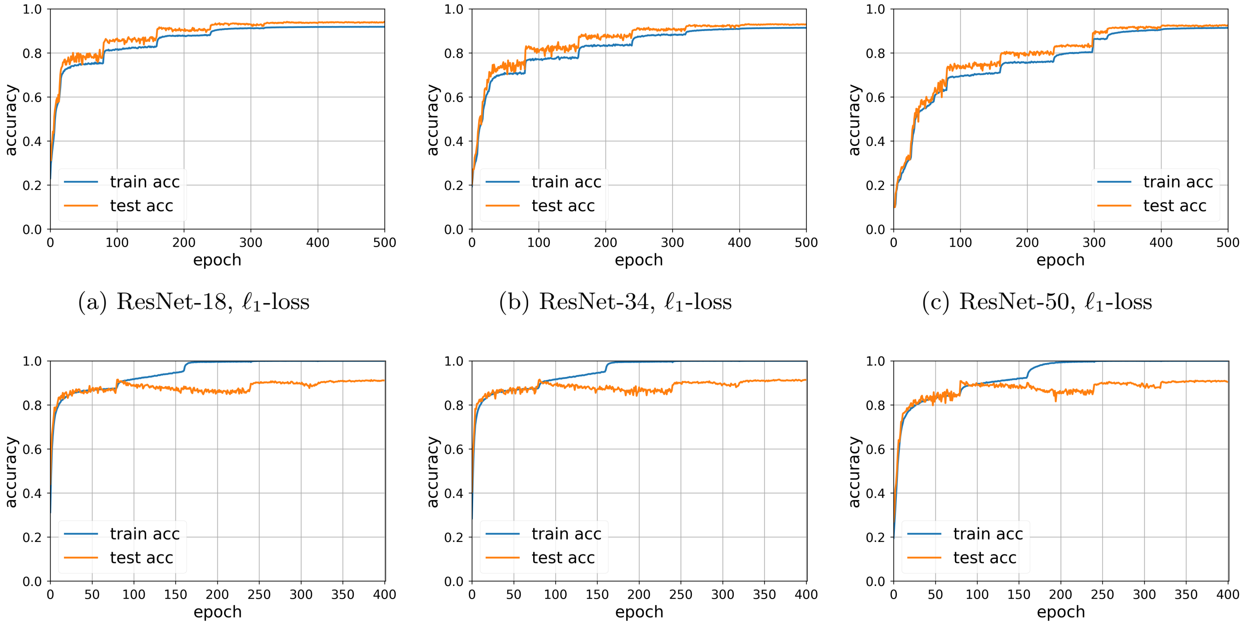
In this section, we provide the training details for the experiments on both CIFAR-10 and CIFAR-100 where 10% of the training data points are randomly labeled. For CIFAR-100 experiment, we use the "loss scaling" trick introduced in [23]. In particular, we denote the neural network by $f_\theta : \mathbb{R}^d \to \mathbb{R}^C$, where d is the input dimension and C is the number of class. The standard ℓ_1-loss for the one-hot encoded label vector can be written (at a single point) as

$$\ell = |f_\theta(x)[c] - 1| + \sum_{c' \neq c} \left| f_\theta(x)[c'] \right|. \tag{12}$$

Here c is the position of the label and $f_\theta(x)[i]$ is the i-th coordinate of the prediction. The rescaled ℓ_1-loss is defined by two parameters k and M as follows:

$$\ell_{\text{scaling}} = k \cdot |f_\theta(x)[c] - M| + \sum_{c' \neq c} \left| f_\theta(x)[c'] \right|. \tag{13}$$

In our simulation, we choose $k = 5$, and $M = 2$. The evolution of the training and test accuracy for CIFAR-10 and CIFAR-100 with both ℓ_1 and CE losses are shown in Figures 5 and 6.



(a) ResNet-18, ℓ_1-loss

(b) ResNet-34, ℓ_1-loss

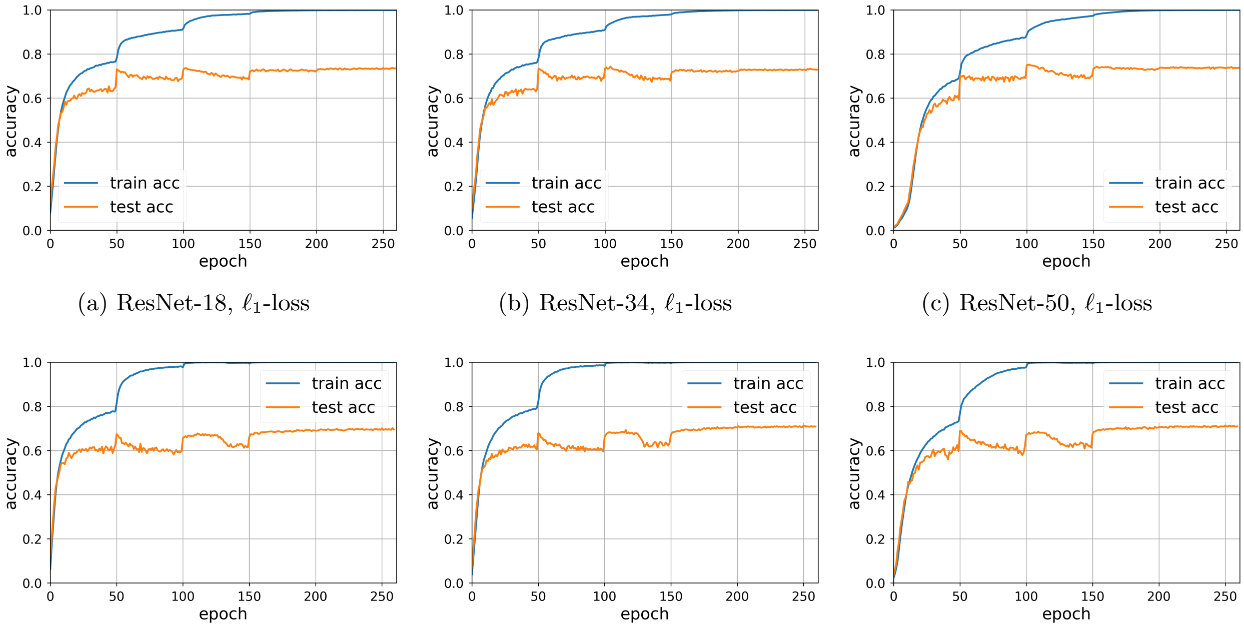
(c) ResNet-50, ℓ_1-loss

(d) ResNet-18, CE

(e) ResNet-34, CE

(f) ResNet-50, CE

Figure 5: We apply ResNet-18, 34, 50 on noisy CIFAR-10 with both ℓ_1-loss and cross-entropy loss (CE). For the training dataset, we randomly choose 10% samples and replace their labels with uniform random labels. We use SGD with initial learning rate $\eta = 0.1$, momentum 0.9, batch size $B = 32$. For every 80 epochs, we decay the learning rate by a factor 0.33. We use standard data augmentation. The initialization is set by default in PyTorch.



(a) ResNet-18, ℓ_1-loss

(b) ResNet-34, ℓ_1-loss

(c) ResNet-50, ℓ_1-loss

(d) ResNet-18, CE

(e) ResNet-34, CE

(f) ResNet-50, CE

Figure 6: We apply ResNet-18, 34, 50 on noisy CIFAR-100 with both ℓ_1-loss and cross-entropy loss (CE). The data generator and optimizer are the same as those in CIFAR-10 experiment. For each trial, we run 260 epochs and decay the learning rate with factor 0.33 for each 50 epochs.

B Proofs of Landscape Analysis

B.1 Proof of Theorem 1

Over-parameterized Regime

We first provide the proof for the 1-layer model. The values of $\mathcal{L}\left(\theta^{\star}\right), \mathcal{L}\left(\theta^{\star}+\Delta\theta\right)$ are provided as

$$\mathcal{L}\left(\theta^{\star}\right)=\frac{1}{m}\sum_{i\in\mathcal{S}}|\epsilon_i|, \quad \mathcal{L}\left(\theta^{\star}+\Delta\theta\right)=\frac{1}{m}\sum_{i\in\bar{\mathcal{S}}}|\langle x_i,\Delta\theta\rangle|+\frac{1}{m}\sum_{i\in\mathcal{S}}|\left\langle x_i,\Delta\theta\right\rangle-\epsilon_i|. \tag{14}$$

Here $\mathcal{S}\in[m]:\{1,2,\ldots,m\}$ is the support of the noise vector $\epsilon=[\epsilon_1,\cdots,\epsilon_m]^\top$, and $\bar{\mathcal{S}}=[m]-\mathcal{S}$. Hence, we have

$$\mathcal{L}\left(\theta^{\star}+\Delta\theta\right)-\mathcal{L}\left(\theta^{\star}\right)=\frac{1}{m}\sum_{i\in\mathcal{S}}\left(|\left\langle x_i,\Delta\theta\right\rangle-\epsilon_i|-|\epsilon_i|\right)+\frac{1}{m}\sum_{i\in\bar{\mathcal{S}}}\left|\langle x_i,\Delta\theta\rangle\right|. \tag{15}$$

Define the subspace

$$U:=\left\{u\in\mathbb{R}^d:\|u\|_\infty\le\gamma,\langle u,x_i\rangle=0,\forall i\in\bar{\mathcal{S}},\ \text{and}\ \left\langle x_i,u\right\rangle\epsilon_i>0,|\left\langle x_i,u\right\rangle|\le|\epsilon_i|,\forall i\in\mathcal{S}\right\}. \tag{16}$$

Note that, since $m\le 0.1d$, the number of constraints in U is upper bounded by $2m\le 0.2d$. Therefore, $\dim(U)\ge 0.8d$. On the other hand, since $U\subset\{u\in\mathbb{R}^d:\|u\|_\infty\le\gamma\}$, we have

$$\inf_{\|\Delta\theta\|_\infty\le\gamma}\left\{\mathcal{L}\left(\theta^{\star}+\Delta\theta\right)-\mathcal{L}\left(\theta^{\star}\right)\right\}\le\inf_{\Delta\theta\in U}\left\{\mathcal{L}\left(\theta^{\star}+\Delta\theta\right)-\mathcal{L}\left(\theta^{\star}\right)\right\}\le\inf_{u\in U}-\frac{1}{m}\sum_{i\in\mathcal{S}}|\left\langle x_i,u\right\rangle|, \tag{17}$$

where the last inequality is due to our definition of the set U. Define the event $\mathcal{E}:=\{$There are at least p_0pm elements of $|\epsilon|$ larger than $t_0\}$. Define $\mathcal{M}_0=\{i:|\epsilon_i|\ge t_0\}$. Using the tail bound of binomial distribution, we have $\mathbb{P}(E)\ge\frac{1}{4}$. Conditioned on the event $\mathcal{E}$, we have $m_0=|\mathcal{M}_0|\ge p_0pm$ and

$$\inf_{u\in U}-\frac{1}{m}\sum_{i\in\mathcal{S}}|\left\langle x_i,u\right\rangle|\le-\sup_{u\in U}\frac{1}{m}\sum_{i\in\mathcal{M}_0}|\left\langle x_i,u\right\rangle|. \tag{18}$$

Hence, it suffices to provide a lower bound for $\sup_{u\in U}\frac{1}{m}\sum_{i\in\mathcal{M}_0}|\left\langle x_i,u\right\rangle|$. To this goal, we define $V:=\left\{u\in\mathbb{R}^d:\|u\|_\infty\le\gamma,\langle u,x_i\rangle=0,\forall i\in\bar{\mathcal{S}},\ \text{and}\ \left\langle x_i,u\right\rangle\epsilon_i>0,\forall i\in\mathcal{S}\right\}$. Hence, we have

$$\begin{aligned}\sup_{u\in U}\frac{1}{m}\sum_{i\in\mathcal{M}_0}|\left\langle x_i,u\right\rangle|&\ge\sup_{u\in V}\frac{1}{m}\sum_{i\in\mathcal{M}_0}|\left\langle x_i,u\right\rangle|\wedge t_0\\&\ge\sup_{u\in V}\frac{1}{m}\sum_{i\in\mathcal{M}_0}|\left\langle x_i,u\right\rangle|\mathbb{1}\left(|\langle x_i,u\rangle|\le t_0\right)\\&\ge\underbrace{\sup_{u\in V}\frac{1}{m}\sum_{i\in\mathcal{M}_0}|\left\langle x_i,u\right\rangle|}_{(A)}-\underbrace{\sup_{u\in V}\frac{1}{m}\sum_{i\in\mathcal{M}_0}|\left\langle x_i,u\right\rangle|\mathbb{1}\left(|\langle x_i,u\rangle|\ge t_0\right)}_{(B)},\end{aligned} \tag{19}$$

where the first inequality follows from the definition of U, V, and $\mathcal{M}_0$. With probability at least $1-\delta$ where $\log(1/\delta) \lesssim d$, we have

$$\begin{aligned}
(A) &\geq \sup_{u \in V} \frac{1}{m} \sum_{i \in \mathcal{M}_0} \langle x_i, u \rangle \\
&= \frac{m_0}{m} \sup_{u \in V} \frac{1}{m_0} \sum_{i \in \mathcal{M}_0} \langle x_i, u \rangle \\
&\overset{(a)}{\geq} \frac{m_0}{m} \left(\mathbb{E} \left[\sup_{u \in V} \frac{1}{m_0} \sum_{i \in \mathcal{M}_0} \langle x_i, u \rangle \right] - \sqrt{d}\gamma \sqrt{\frac{\log(1/\delta)}{m_0}} \right) \\
&\overset{(b)}{\gtrsim} \frac{m_0}{m} \sqrt{d}\gamma \left(\sqrt{\frac{d}{m_0}} - \sqrt{\frac{\log(1/\delta)}{m_0}} \right) \\
&\gtrsim \sqrt{\frac{p_0 p}{m}} d\gamma.
\end{aligned} \tag{20}$$

Here in (a), we used Theorem 3.1 in [40], and in (b) we used Sudakov's inequality. To provide a bound for (B), we first use Hölder's inequality to write

$$\begin{aligned}
(B) &\leq \sup_{u \in V} \frac{1}{m} \left(\sum_{i \in \mathcal{M}_0} |\langle x_i, u \rangle| \right) \sup_{i \in \mathcal{M}_0} \mathbb{1} \left(|\langle x_i, u \rangle| \geq t_0 \right) \\
&\leq \sup_{u \in V} \frac{1}{m} \left(\sum_{i \in \mathcal{M}_0} |\langle x_i, u \rangle| \right) \sup_{i \in \mathcal{M}_0} \mathbb{1} \left(\|x_i\| \gamma \geq t_0 \right).
\end{aligned} \tag{21}$$

For the second part, we have

$$\begin{aligned}
\mathbb{P} \left(\sup_{i \in \mathcal{M}_0} \mathbb{1} \left(\|x_i\| \gamma \geq t_0 \right) = 1 \right) &\leq \mathbb{P} \left(\mathbb{1} \left(\sup_{i \in \mathcal{M}_0} \|x_i\| \gamma \geq t_0 \right) = 1 \right) \\
&= \mathbb{P} \left(\sup_{i \in \mathcal{M}_0} \|x_i\| \gamma \geq t_0 \right) \\
&\leq m_0 \mathbb{P} \left(\|x_i\| \gamma \geq t_0 \right) \\
&\lesssim e^{\log(m_0) - C\left(\frac{t_0}{\gamma}\right)^2} \\
&\leq e^{-\Omega(d)}.
\end{aligned} \tag{22}$$

provided that $\gamma \lesssim \frac{t_0}{\sqrt{d}} \wedge 1$. Hence, we have

$$(B) = 0, \qquad \text{with probability of } e^{-\Omega(d)}. \tag{23}$$

Therefore, combining (A) and (B) with the choice of $\delta = \frac{1}{2}$, we have

$$\inf_{\|\Delta\theta\|_\infty \leq \gamma} \left\{ \mathcal{L}\left(\theta^\star + \Delta\theta\right) - \mathcal{L}\left(\theta^\star\right) \right\} \lesssim -\sqrt{\frac{p_0 p}{m}} d\gamma, \tag{24}$$

with probability of at least $\frac{1}{16}$.

For general N-layer models, we consider the true solution $\mathbf{w} \in \mathcal{W}$ with $w_1 = \theta^\star$, and $w_2 = \cdots w_N = \mathbf{1}$. Moreover, for $\Delta\mathbf{w}$ we choose $\Delta w_2 = \cdots \Delta w_N = \mathbf{0}$. It is easy to verify that

$$\mathcal{L}(\mathbf{w}) - \mathcal{L}(\mathbf{w}+\Delta\mathbf{w}) = \frac{1}{m}\sum_{i\in\mathcal{S}} \left(\left|\langle x_i, \Delta w_1\rangle - \epsilon_i\right| - |\epsilon_i|\right) + \frac{1}{m}\sum_{i\in\bar{\mathcal{S}}} \left|\langle x_i, \Delta w_1\rangle\right|. \tag{25}$$

Therefore, an argument similar to 1-layer model can be used to write

$$\inf_{\mathbf{w}\in\mathcal{W}} \inf_{\mathbf{w}':\|\mathbf{w}-\mathbf{w}'\|_\infty\le\gamma} \left\{\mathcal{L}(\mathbf{w}) - \mathcal{L}(\mathbf{w}')\right\} \le \inf_{\|\Delta w_1\|_\infty\le\gamma} \left\{\mathcal{L}(\mathbf{w}) - \mathcal{L}(\mathbf{w}+\Delta\mathbf{w})\right\} \lesssim -\sqrt{\frac{p_0 p}{m}} d\gamma, \tag{26}$$

with probability of $\frac{1}{16}$.

Under-parameterized Regime

Given $\mathbf{w} \in \mathcal{W}$ and any $\Delta\mathbf{w}$, consider $\mathbf{w}' = \mathbf{w} + \Delta\mathbf{w}$ and define

$$\Delta\theta = \sum_{i=1}^{N} (\theta^\star)^{\frac{N-i}{N}} \odot \sum_{j_1,\cdots,j_i} \Delta w_{j_1} \odot \cdots \odot \Delta w_{j_i}. \tag{27}$$

We have

$$\begin{aligned}\mathcal{L}(\mathbf{w}) - \mathcal{L}(\mathbf{w}+\Delta\mathbf{w}) &= \frac{1}{m}\sum_{i\in\mathcal{S}} \left(\left|\langle x_i, \Delta\theta\rangle - \epsilon_i\right| - |\epsilon_i|\right) + \frac{1}{m}\sum_{i\in\bar{\mathcal{S}}} \left|\langle x_i, \Delta\theta\rangle\right| \\ &\ge \frac{1}{m}\sum_{i\in\bar{\mathcal{S}}} \left|\langle x_i, \Delta\theta\rangle\right| - \frac{1}{m}\sum_{i\in\mathcal{S}} \left|\langle x_i, \Delta\theta\rangle\right|.\end{aligned} \tag{28}$$

Hence, it suffices to show that, with probability of $1 - e^{-\Omega(d)}$,

$$\inf_{\Delta\theta\in\mathbf{R}^d} \left\{\frac{1}{m}\sum_{i\in\bar{\mathcal{S}}} \left|\langle x_i, \Delta\theta\rangle\right| - \frac{1}{m}\sum_{i\in\mathcal{S}} \left|\langle x_i, \Delta\theta\rangle\right|\right\} \ge 0. \tag{29}$$

Note that the above inequality is invariant with respect to scaling. Hence, it suffices to show that it holds for arbitrary $\Delta\theta \in \mathbb{S}^{d-1}$ where $\mathbb{S}^{d-1} := \{x \in \mathbb{R}^d : \|x\| = 1\}$ is the standard sphere. Hence, it suffices to show

$$\begin{aligned}&\inf_{\Delta\theta\in\mathbb{S}^{d-1}} \left\{\frac{1}{m}\sum_{i\in\bar{\mathcal{S}}} \left|\langle x_i, \Delta\theta\rangle\right| - \frac{1}{m}\sum_{i\in\mathcal{S}} \left|\langle x_i, \Delta\theta\rangle\right|\right\} \\ &\ge \inf_{\Delta\theta\in\mathbb{S}^{d-1}} \left\{\frac{1}{m}\sum_{i\in\bar{\mathcal{S}}} \left|\langle x_i, \Delta\theta\rangle\right|\right\} - \sup_{\Delta\theta\in\mathbb{S}^{d-1}} \left\{\frac{1}{m}\sum_{i\in\mathcal{S}} \left|\langle x_i, \Delta\theta\rangle\right|\right\} \ge 0.\end{aligned} \tag{30}$$

For the first term, applying Lemma 6, we have that with probability at least $1 - e^{-\Omega(d)}$

$$\inf_{\Delta\theta\in\mathbb{S}^{d-1}} \frac{1}{m}\sum_{i\in\bar{\mathcal{S}}} \left|\langle x_i, \Delta\theta\rangle\right| \ge \sqrt{\frac{2}{\pi}}(1-p) - \sqrt{\frac{(1-p)d}{m}}. \tag{31}$$

Similarly, for the second part, with probability of at least $1-e^{-\Omega(d)}$, we have

$$\sup_{\Delta\theta\in\mathbb{S}^{d-1}}\frac{1}{m}\sum_{i\in\mathcal{S}}|\langle x_i,\Delta\theta\rangle|\leq\sqrt{\frac{2}{\pi}}p+\sqrt{\frac{pd}{m}}. \tag{32}$$

Combining both parts, we have that with probability at least $1-e^{-\Omega(d)}$

$$\inf_{\Delta\theta\in\mathbb{S}^{d-1}}\left\{\frac{1}{m}\sum_{i\in\bar{\mathcal{S}}}|\langle x_i,\Delta\theta\rangle|\right\}-\sup_{\Delta\theta\in\mathbb{S}^{d-1}}\left\{\frac{1}{m}\sum_{i\in\mathcal{S}}|\langle x_i,\Delta\theta\rangle|\right\}\geq\sqrt{\frac{2}{\pi}}(1-2p)-2\sqrt{\frac{d}{m}}\geq 0. \tag{33}$$

The last inequality follows from the fact that $m\gtrsim\frac{d}{(1-2p)^2}$. This completes the proof. □

B.2 Proof of Theorem 4

Given any $\Delta\mathbf{w}=[\Delta w_1,\cdots,\Delta w_N]^\top$, the following equality holds for any point $\mathbf{w}=\mathbf{w}^\star+\Delta\mathbf{w}$ where $\mathbf{w}^\star=[\sqrt[N]{\theta^\star},\cdots,\sqrt[N]{\theta^\star}]^\top$:

$$\begin{aligned}(\sqrt[N]{\theta^\star}+\Delta w_1)\odot\cdots\odot(\sqrt[N]{\theta^\star}+\Delta w_N)-\theta^\star=&\underbrace{\sum_{i=1}^{N-1}(\theta^\star)^{\frac{N-i}{N}}\odot\sum_{j_1,\cdots,j_i}\Delta w_{j_1}\odot\cdots\odot\Delta w_{j_i}}_{:=\Delta\theta_1}\\&+\underbrace{\Delta w_1\odot\cdots\odot\Delta w_N}_{:=\Delta\theta_2}.\end{aligned} \tag{34}$$

Hence, we have

$$\mathcal{L}(\mathbf{w}^\star)-\mathcal{L}(\mathbf{w}^\star+\Delta\mathbf{w})=\frac{1}{m}\sum_{i\in\mathcal{S}}\left(|\langle x_i,\Delta\theta_1+\Delta\theta_2\rangle-\epsilon_i|-|\epsilon_i|\right)+\frac{1}{m}\sum_{i\in\bar{\mathcal{S}}}|\langle x_i,\Delta\theta_1+\Delta\theta_2\rangle|. \tag{35}$$

For simplicity, we denote Θ_1 as the set of $\Delta\theta_1$ defined in (35) with $\|\Delta\mathbf{w}\|_\infty\leq\gamma$. Similarly, Θ_2 is the set of $\Delta\theta_2$ defined in (35) with $\|\Delta\mathbf{w}\|_\infty\leq\gamma$.

Lower bound. To prove the lower bound, one can write

$$\begin{aligned}\mathcal{L}(\mathbf{w}^\star)-\mathcal{L}(\mathbf{w}^\star+\Delta\mathbf{w})&\geq\frac{1}{m}\sum_{i\in\bar{\mathcal{S}}}|\langle x_i,\Delta\theta_1+\Delta\theta_2\rangle|-\frac{1}{m}\sum_{i\in\mathcal{S}}|\langle x_i,\Delta\theta_1+\Delta\theta_2\rangle|\\&\geq\frac{1}{m}\sum_{i\in\bar{\mathcal{S}}}|\langle x_i,\Delta\theta_1\rangle|-\frac{1}{m}\sum_{i\in\mathcal{S}}|\langle x_i,\Delta\theta_1\rangle|-\frac{1}{m}\sum_{i=1}^{m}|\langle x_i,\Delta\theta_2\rangle|\end{aligned} \tag{36}$$

Hence, we have

$$\begin{aligned}\inf_{\|\Delta\mathbf{w}\|_\infty\leq\gamma}\{\mathcal{L}(\mathbf{w}^\star)-\mathcal{L}(\mathbf{w}^\star+\Delta\mathbf{w})\}\geq&\inf_{\Delta\theta_1\in\Theta_1}\left\{\frac{1}{m}\sum_{i\in\bar{\mathcal{S}}}|\langle x_i,\Delta\theta_1\rangle|-\frac{1}{m}\sum_{i\in\mathcal{S}}|\langle x_i,\Delta\theta_1\rangle|\right\}\\&-\sup_{\Delta\theta_2\in\Theta_2}\left\{\frac{1}{m}\sum_{i=1}^{m}|\langle x_i,\Delta\theta_2\rangle|\right\}.\end{aligned} \tag{37}$$

First, we bound the term $\inf_{\Delta\theta_1\in\Theta_1}\left\{\frac{1}{m}\sum_{i\in\bar{\mathcal{S}}}|\langle x_i,\Delta\theta_1\rangle| - \frac{1}{m}\sum_{i\in\mathcal{S}}|\langle x_i,\Delta\theta_1\rangle|\right\}$. It is easy to see that the vector $\Delta\theta_1$ is k-sparse and has the same sparsity pattern as $\theta^\star$. Therefore, according to Lemma 6, there exist universal constants $C, c > 0$ such that the following hold

$$\mathbb{P}\left(\sup_{\Delta\theta_1\in\Theta_1}\left|\frac{1}{m'\left\|\Delta\theta_1\right\|}\sum_{i\in\bar{\mathcal{S}}}|\left\langle x_i,\Delta\theta_1\right\rangle| - \sqrt{\frac{2}{\pi}}\right| \geq C\sqrt{\frac{k}{m'}}+\delta\right) \leq e^{-cm'\delta^2}, \tag{38}$$

and

$$\mathbb{P}\left(\sup_{\Delta\theta_1\in\Theta_1}\left|\frac{1}{m''\left\|\Delta\theta_1\right\|}\sum_{i\in\mathcal{S}}|\left\langle x_i,\Delta\theta_1\right\rangle| - \sqrt{\frac{2}{\pi}}\right| \geq C\sqrt{\frac{k}{m''}}+\delta\right) \leq e^{-cm''\delta^2}. \tag{39}$$

Here $m' = (1-p)m$, and $m'' = pm$. The inequality (38) implies that

$$\mathbb{P}\left(\frac{1}{m}\sum_{i\in\bar{\mathcal{S}}}|\langle x_i,\Delta\theta_1\rangle| \geq \|\Delta\theta_1\|\left(\sqrt{\frac{2}{\pi}}(1-p) - C\sqrt{\frac{(1-p)k}{m}} - \delta_1\right), \forall\Delta\theta_1\in\Theta_1\right) \geq 1 - e^{-\frac{cm\delta_1^2}{1-p}}. \tag{40}$$

Similarly, the inequality (39) leads to

$$\mathbb{P}\left(\frac{1}{m}\sum_{i\in\mathcal{S}}|\langle x_i,\Delta\theta_1\rangle| \leq \|\Delta\theta_1\|\left(\sqrt{\frac{2}{\pi}}p + C\sqrt{\frac{pk}{m}} + \delta_2\right), \forall\Delta\theta_1\in\Theta_1\right) \geq 1 - e^{-\frac{cm\delta_2^2}{p}}. \tag{41}$$

Upon setting $\delta_1 = \sqrt{\frac{(1-p)k}{m}}$ and $\delta_2 = \sqrt{\frac{pk}{m}}$, with probability of $1 - e^{-\Omega(k)}$, for all $\Delta\theta_1\in\Theta_1$, we have

$$\frac{1}{m}\sum_{i\in\bar{\mathcal{S}}}|\langle x_i,\Delta\theta_1\rangle| - \frac{1}{m}\sum_{i\in\mathcal{S}}|\langle x_i,\Delta\theta_1\rangle| \geq \sqrt{\frac{2}{\pi}}(1-2p) - C'\sqrt{\frac{k}{m}} \geq 0. \tag{42}$$

In the last inequality we used the assumption $m \gtrsim \frac{k}{(1-2p)^2}$. The above argument implies

$$\inf_{\Delta\theta_1\in\Theta_1}\left\{\frac{1}{m}\sum_{i\in\bar{\mathcal{S}}}|\langle x_i,\Delta\theta_1\rangle| - \frac{1}{m}\sum_{i\in\mathcal{S}}|\langle x_i,\Delta\theta_1\rangle|\right\} \geq 0, \tag{43}$$

with probability of at least $1 - e^{-\Omega(k)}$. Now we turn to bound the second part $\sup_{\Delta\theta_2\in\Theta_2}\left\{\frac{1}{m}\sum_{i=1}^m|\left\langle x_i,\Delta\theta_2\right\rangle|\right\}$. To this goal, we first apply Lemma 6, which leads to

$$\mathbb{P}\left(\sup_{\Delta\theta_2\in\Theta_2}\left|\frac{1}{m\left\|\Delta\theta_2\right\|}\sum_{i=1}^m|\left\langle x_i,\Delta\theta_2\right\rangle| - \sqrt{\frac{2}{\pi}}\right| \geq C\sqrt{\frac{d}{m}}+\delta\right) \leq e^{-cm\delta^2}. \tag{44}$$

Therefore, upon setting $\delta = \sqrt{\frac{d}{m}}$, with probability of $1 - e^{-\Omega(d)}$, we have

$$\begin{aligned}
\sup_{\Delta\theta_2 \in \Theta_2} \left\{ \frac{1}{m} \sum_{i=1}^{m} |\left\langle x_i, \Delta\theta_2 \right\rangle| \right\} &\leq \sup_{\Delta\theta_2 \in \Theta_2} \|\Delta\theta_2\| \left(\sqrt{\frac{2}{\pi}} + (C+1)\sqrt{\frac{d}{m}} \right) \\
&\lesssim \sqrt{\frac{d}{m}} \sup_{\|\Delta\mathbf{w}\|_\infty \leq \gamma} \|\Delta w_1 \odot \cdots \odot \Delta w_N\| \\
&= \sqrt{\frac{d}{m}} \sup_{|\Delta w_i[j]| \leq \gamma} \sqrt{\sum_{j \in [d]} \left(\prod_{i \in [N]} \Delta w_i[j] \right)^2} \\
&= \frac{d}{\sqrt{m}} \gamma^N.
\end{aligned} \tag{45}$$

Here $\Delta w_i[j]$ is the j-th element of Δw_i. Therefore, we conclude that

$$\inf_{\|\Delta\mathbf{w}\|_\infty \leq \gamma} \left\{ \mathcal{L}(\mathbf{w}^\star) - \mathcal{L}\left(\mathbf{w}^\star + \Delta\mathbf{w}\right) \right\} \gtrsim -\frac{d}{\sqrt{m}} \gamma^N, \tag{46}$$

with probability of $1 - e^{-\Omega(k)}$, thereby completing the proof of the lower bound.

Upper bound. To this goal, we first define a restricted set of perturbation vectors

$$\mathcal{V} := \left\{ \mathbf{v} : \|\mathbf{v}\|_\infty \leq \gamma, v_i[j] = 0, \forall i \in [N], j \in \mathrm{supp}(\theta^\star) \right\}, \tag{47}$$

where $\mathrm{supp}(\theta^\star)$ is the support of $\theta^\star$. Based on this definition, we have

$$\begin{aligned}
\inf_{\|\Delta\mathbf{w}\|_\infty \leq \gamma} \left\{ \mathcal{L}(\mathbf{w}^\star) - \mathcal{L}\left(\mathbf{w}^\star + \Delta\mathbf{w}\right) \right\} &\leq \inf_{\Delta\mathbf{w} \in \mathcal{V}} \left\{ \mathcal{L}(\mathbf{w}^\star) - \mathcal{L}\left(\mathbf{w}^\star + \Delta\mathbf{w}\right) \right\} \\
&= \inf_{\Delta\mathbf{w} \in \mathcal{V}} \left\{ \frac{1}{m} \sum_{i \in \mathcal{S}} \left(|\left\langle x_i, \Delta\theta_2 \right\rangle - \epsilon_i| - |\epsilon_i| \right) + \frac{1}{m} \sum_{i \in \bar{\mathcal{S}}} |\langle x_i, \Delta\theta_2 \rangle| \right\}.
\end{aligned} \tag{48}$$

where $\Delta\theta_2 = \Delta w_1 \odot \cdots \odot \Delta w_N$ is the same as before. Note that for any $\|\Delta\mathbf{w}\|_\infty \leq \gamma$, we have $\|\Delta\theta_2\| \leq \sqrt{d}\gamma^N$. Moreover, this bound is attainable when $\Delta w_i \equiv [\pm\gamma, \cdots, \pm\gamma]^\top$. Hence, we have

$$\begin{aligned}
&\inf_{\Delta\mathbf{w} \in \mathcal{V}} \left\{ \frac{1}{m} \sum_{i \in \mathcal{S}} \left(|\left\langle x_i, \Delta\theta_2 \right\rangle - \epsilon_i| - |\epsilon_i| \right) + \frac{1}{m} \sum_{i \in \bar{\mathcal{S}}} |\langle x_i, \Delta\theta_2 \rangle| \right\} \\
&\leq \inf_{\|\Delta\theta_2\| \leq \sqrt{d}\gamma^N} \left\{ \frac{1}{m} \sum_{i \in \mathcal{S}} \left(|\left\langle x_i, \Delta\theta_2 \right\rangle - \epsilon_i| - |\epsilon_i| \right) + \frac{1}{m} \sum_{i \in \bar{\mathcal{S}}} |\langle x_i, \Delta\theta_2 \rangle| \right\}.
\end{aligned} \tag{49}$$

An argument similar to the proof of Theorem 1 can be used to show that, with probability of at least 1/16, we have

$$\inf_{\|\Delta\theta_2\| \leq \sqrt{d}\gamma^N} \left\{ \frac{1}{m} \sum_{i \in S} \left(|\left\langle x_i, \Delta\theta_2 \right\rangle - \epsilon_i| - |\epsilon_i| \right) + \frac{1}{m} \sum_{i \in \bar{S}} |\langle x_i, \Delta\theta_2 \rangle| \right\} \lesssim -\sqrt{\frac{p_0 p(d-k)}{m}} \sqrt{d}\gamma^N. \tag{50}$$

Recalling that $k \ll d$, we have with probability of $1/16$

$$\inf_{\|\Delta \mathbf{w}\|_\infty \le \gamma} \left\{\mathcal{L}(\mathbf{w}^\star) - \mathcal{L}\left(\mathbf{w}^\star + \Delta \mathbf{w}\right)\right\} \lesssim -\sqrt{p_0 p}\frac{d}{\sqrt{m}}\gamma^N. \tag{51}$$

This completes the proof. □

C Proofs of Convergence Analysis

C.1 Proof of Theorem 2

For simplicity of notation, we denote $u = w_1$ and $v = w_2$. Moreover, without loss of generality, we assume that the elements of $\theta^\star$ are arranged in descending order, i.e., $\theta_1^\star \ge \cdots \ge \theta_k^\star > \theta_{k+1}^\star = \cdots = \theta_d^\star = 0$, and the initial point satisfies $u_i, v_i = \Theta\left(\sqrt{\alpha}\right), \forall i \in [d]$. Moreover, for $v \in \mathbf{R}^d$, we define $v_{:i} = [v_1, \cdots, v_i]^\top$ and $v_{i:} = [v_i, \cdots, v_d]^\top$. For short, we denote $v_{-i} = v_{i+1:}$. Finally, we define $\kappa = \theta_1^\star/\theta_k^\star$ as the condition number. Moreover, without loss of generality, we assume that $\theta_1^\star \ge 1 \ge \theta_k^\star$.

First, according to Proposition 1, the sub-differential of $\mathcal{L}(u, v)$ is uniformly concentrated around its population gradient. In particular, with probability at least $1 - Ce^{-cm\delta^2}$, we have

$$u_{i,t+1} = u_{i,t} + \eta\frac{\theta_i^\star - u_{i,t}v_{i,t}}{\|u_t \odot v_t - \theta^\star\|}v_{i,t} + \eta\delta_i v_{i,t}, \text{ and } |\delta_i| \le \delta, \forall i \in [d], \tag{52}$$

$$v_{i,t+1} = v_{i,t} + \eta\frac{\theta_i^\star - u_{i,t}v_{i,t}}{\|u_t \odot v_t - \theta^\star\|}u_{i,t} + \eta\delta_i u_{i,t}, \text{ and } |\delta_i| \le \delta, \forall i \in [d]. \tag{53}$$

Hence, we have

$$u_{i,t+1}v_{i,t+1} = u_{i,t}v_{i,t} + \eta\left(\frac{\theta_i^\star - u_{i,t}v_{i,t}}{\|u_t \odot v_t - \theta^\star\|} + \delta_i\right)(u_{i,t}^2 + v_{i,t}^2) + \eta^2\left(\frac{\theta_i^\star - u_{i,t}v_{i,t}}{\|u_t \odot v_t - \theta^\star\|} + \delta_i\right)^2 u_{i,t}v_{i,t}. \tag{54}$$

Moreover, we have

$$u_{i,t+1}^2 + v_{i,t+1}^2 = \left(u_{i,t}^2 + v_{i,t}^2\right)\left(1 + \eta^2\left(\frac{(\theta_i^\star - u_{i,t}v_{i,t})}{\|u_t \odot v_t - \theta^\star\|} + \delta_i\right)^2\right) + 4\eta\left(\frac{\theta_i^\star - u_{i,t}v_{i,t}}{\|u_t \odot v_t - \theta^\star\|} + \delta_i\right)u_{i,t}v_{i,t}. \tag{55}$$

Signal Dynamics

We first study the behavior of the signal term $S_t = u_{:k,t}v_{:k,t}$ for the first k components of the model $u_t \odot v_t$. We divide the dynamics into $k+1$ stages. In the first k stages, each component $u_{i,t}v_{i,t}$ converges to $\theta_i^\star$ sequentially. Once all the components are close to the ground truth, the distance between signal term and the ground truth $\|S_t - \theta_{:k}^\star\|$ will further decrease to $\mathcal{O}\left(\sqrt{d^2 m\alpha^2} \vee \eta\theta_1^\star\right)$.

Stage 1: In this stage, the first component $u_1 v_1$ grows to $\theta_1 - \delta \left\|\theta^\star\right\|$ within $\Theta\left(\frac{\|\theta^\star\|}{\eta\theta_1^\star}\log\left(\frac{1}{\alpha}\right)\right)$ iterations. At the initial point, we have $u_{1,0}v_{1,0} = \Theta(\alpha)$. For iteration $t+1$, according to (54), we have

$$\begin{aligned} u_{1,t+1}v_{1,t+1} &\geq u_{1,t}v_{1,t} + \eta\left(\frac{\theta_1^\star - u_{1,t}v_{1,t}}{\|u_t \odot v_t - \theta^\star\|} + \delta_1\right)\left(u_{1,t}^2 + v_{1,t}^2\right) \\ &\geq u_{1,t}v_{1,t} + 2\eta\left(\frac{\theta_1^\star - u_{1,t}v_{1,t}}{\|u_t \odot v_t - \theta^\star\|} + \delta_1\right)u_{1,t}v_{1,t}. \end{aligned} \tag{56}$$

We further divide our analysis into two substages. In the first substage, we have $u_1 v_1 \leq \theta_1^\star/2$. Note that $\|u_t \odot v_t - \theta^\star\| = \mathcal{O}(\|\theta^\star\|)$ and $|\delta_1| \leq \delta \lesssim \frac{1}{\kappa}$. Hence, (56) can be further simplified as

$$u_{1,t+1}v_{1,t+1} \geq \left(1 + \frac{1}{4}\frac{\eta\theta_1^\star}{\|\theta^\star\|}\right)u_{1,t}v_{1,t}. \tag{57}$$

Therefore, this stage ends within $\mathcal{O}\left(\frac{\|\theta^\star\|}{\eta\theta_1^\star}\log\left(\frac{1}{\alpha}\right)\right)$ iterations. In the second substage, we have $u_1 v_1 \geq \theta_1^\star/2$. Upon defining $x_t = \theta_1^\star - u_{1,t}v_{1,t}$, one can write

$$\begin{aligned} x_{t+1} &\leq \left(1 - 2\eta\frac{u_{1,t}v_{1,t}}{\|u_t \odot v_t - \theta^\star\|}\right)x_t - 2\eta\delta_1 u_t \odot v_t \\ &\leq \left(1 - \frac{\eta\theta_1^\star}{\|\theta^\star\|}\right)x_t + \eta\delta\theta_1^\star. \end{aligned} \tag{58}$$

Hence, within additional $\mathcal{O}\left(\|\theta^\star\|/(\eta\theta_1^\star)\right)$ iterations, we have $u_{1,T_1}v_{1,T_1} = \theta_1^\star \pm \delta\left\|\theta^\star\right\|$. Overall, within $T_1 = \mathcal{O}\left(\frac{\|\theta^\star\|}{\eta\theta_1^\star}\log\left(\frac{1}{\alpha}\right)\right)$ iterations, we have $u_{1,T_1}v_{1,T_1} = \theta_1^\star \pm \delta\left\|\theta^\star\right\|$. Now, we turn to show the lower bound on T_1 by analyzing the trajectory of $u_{1,t}^2 + v_{1,t}^2$. Due to (55), when $u_{1,t}^2 + v_{1,t}^2 \leq \frac{\theta_1^\star}{2}$, we have

$$u_{1,t+1}^2 + v_{1,t+1}^2 \leq \left(u_{1,t}^2 + v_{1,t}^2\right)\left(1 + 10\frac{\eta\theta_1^\star}{\|\theta^\star\|}\right). \tag{59}$$

Hence, at least $\Omega\left(\frac{\|\theta^\star\|}{\eta\theta_1^\star}\log\left(\frac{1}{\alpha}\right)\right)$ iterations are needed for $u_{1,t}^2 + v_{1,t}^2$ to be larger than $\frac{\theta_1^\star}{2}$. Since $u_{1,t}^2 + v_{1,t}^2 \geq u_{1,t}v_{1,t}$, we immediately obtain $T_1 = \Omega\left(\frac{\|\theta^\star\|}{\eta\theta_1^\star}\log\left(\frac{1}{\alpha}\right)\right)$.

Stage 2 to Stage *k*: In the next $k-1$ stages, each component $u_i v_i$ will converge to $\theta_i^\star \pm \delta\left\|\theta^\star\right\|$ sequentially. To show this, we use an inductive argument. In each stage i, we assume that the first $i-1$ components have already converged close to $\theta_j^\star, \forall j \in [i-1]$. Hence, we have $\|u_t \odot v_t - \theta^\star\| = \Theta\left(\left\|\theta_{-(i-1)}^\star\right\|\right)$. Repeating the procedure in Stage 1, we can show that, at stage i, $T_i = \mathcal{O}\left(\frac{\left\|\theta_{-(i-1)}^\star\right\|}{\eta\theta_i^\star}\log\left(\frac{1}{\alpha}\right)\right)$ iterations are needed for $u_i v_i$ to converge to $\theta_i^\star \pm \delta\left\|\theta^\star\right\|$. Overall, after $T = T_1 + \cdots T_k = \mathcal{O}\left(\frac{k^{3/2}}{\eta}\log\left(\frac{1}{\alpha}\right)\right)$ iterations, we have

$$\left\|u_{:k,T}v_{:k,T} - \theta_{:k,T}^\star\right\| \lesssim \sqrt{k}\delta\left\|\theta^\star\right\|. \tag{60}$$

Stage $k+1$: In the final stage, the signal term will quickly decrease to $\mathcal{O}\left(\sqrt{d^2 m}\alpha^{1-\Theta(\delta)} \vee \eta\theta_1^\star\right)$ within $T_{k+1} = \mathcal{O}\left(\frac{\delta\|\theta^\star\|}{\eta\theta_k^\star}\right)$ iterations. To show this, we write

$$\theta_{:k}^\star - S_{t+1} = \theta_{:k}^\star - S_t - \eta\frac{u_{:k,t}^2 + v_{:k,t}^2}{\|u_t \odot v_t - \theta^\star\|} \odot (\theta_{:k}^\star - S_t) - \eta\boldsymbol{\delta}\left(u_t^2 + v_t^2\right) + \eta^2\left(\frac{\theta_{:k}^\star - S_t}{\|u_t \odot v_t - \theta^\star\|} + \boldsymbol{\delta}\right)^2 \odot S_t. \tag{61}$$

Here we denote $\boldsymbol{\delta} = [\delta_1, \cdots, \delta_k]^\top$. Note that $\|u_t \odot v_t - \theta^\star\| \le \|\theta_{:k}^\star - S_t\| + \|E_t\|$. Moreover, based on our assumption, we have $\|E_t\| \lesssim \sqrt{d}\alpha^{1-\Theta(\delta)}$. Finally, the balanced property implies that $|u_{i,t} - v_{i,t}| = \mathcal{O}\left(\alpha^{1-\Theta(\delta)}\right)$ (this will be proven later). Hence, we have

$$\begin{aligned}
\|\theta_{:k}^\star - S_{t+1}\| &\le \left\|(\theta_{:k}^\star - S_t) \odot \left(\mathbf{1} - \eta\frac{u_t^2 + v_t^2}{\|u_t \odot v_t - \theta^\star\|} + \eta^2\frac{(\theta_{:k}^\star - S_t) \odot S_t}{\|u_t \odot v_t - \theta^\star\|^2}\right)\right\| + 4\eta\delta\|\theta^\star\| \\
&\le \|\theta_{:k}^\star - S_t\|\left(1 - \frac{\eta}{2}\frac{\theta_k^\star}{\|\theta_{:k}^\star - S_t\| + \|E_t\|}\right) + 4\eta\delta\|\theta^\star\| \\
&\le \|\theta_{:k}^\star - S_t\| - 0.25\eta\theta_k^\star + 4\eta\delta\|\theta^\star\| \\
&\le \|\theta_{:k}^\star - S_t\| - 0.1\eta\theta_k^\star.
\end{aligned} \tag{62}$$

Here, we used the fact that $\delta \lesssim \frac{1}{\kappa}$. On the other hand, we have

$$\begin{aligned}
\|S_{t+1} - S_t\| &\le \frac{\eta}{\|u_t \odot v_t - \theta^\star\|}\left\|\left(u_t^2 + v_t^2\right) \odot (\theta_{:k}^\star - S_t)\right\| + 4\eta\delta\|\theta^\star\| \\
&\overset{(a)}{\le} 2\eta\theta_1^\star + 4\eta\delta\|\theta^\star\| \\
&\le 3\eta\theta_1^\star,
\end{aligned} \tag{63}$$

where in (a) we used Lemma 7. The above inequality indicates that the signal propagation in each step is upper bounded by $\mathcal{O}(\eta\theta_1^\star)$. Hence, we conclude that within $T_{k+1} = \mathcal{O}\left(\frac{\delta\|\theta^\star\|}{\eta\theta_k^\star}\right)$ iterations, we have $\|\theta_{:k}^\star - S_t\| \lesssim \sqrt{d^2 m}\alpha^{1-\Theta(\delta)} \vee \eta\theta_1^\star$. Since $\delta \lesssim \frac{1}{\kappa}$, the total iteration complexity is upper bounded by $T' = T_1 + \cdots T_{k+1} = \mathcal{O}\left(\frac{k^{3/2}}{\eta}\log\left(\frac{1}{\alpha}\right)\right)$.

Residual Dynamics

Now, we analyze the residual dynamics. Instead of analyzing the dynamics of $u_{i,t}v_{i,t}$, we analyze its surrogate $u_{i,t}^2 + v_{i,t}^2$. Based on (55), we can naturally bound it as follows

$$u_{i,t+1}^2 + v_{i,t+1}^2 \le u_{i,t}^2 + v_{i,t}^2 + 6\eta\delta u_{i,t}v_{i,t} \le (1 + 3\eta\delta)\left(u_{i,t}^2 + v_{i,t}^2\right). \tag{64}$$

Therefore, during the training process, we can bound the residual term as

$$u_{i,t}^2 + v_{i,t}^2 \lesssim \alpha(1 + \eta\delta)^{\mathcal{O}\left(\frac{k^{3/2}}{\eta}\log\left(\frac{1}{\alpha}\right)\right)} \lesssim \alpha^{1-\mathcal{O}\left(k^{3/2}\delta\right)}. \tag{65}$$

Hence, we have

$$\|E_t\| \le \left(\sum_{i=k+1}^{d} u_{i,t}^2 + v_{i,t}^2\right)^{1/2} \lesssim \sqrt{d}\alpha^{1-\mathcal{O}\left(k^{3/2}\delta\right)}. \tag{66}$$

Therefore, we conclude that within $\bar{T} = \mathcal{O}\left(\frac{k^{3/2}}{\eta}\log\left(\frac{1}{\alpha}\right)\right)$ iterations, we have

$$\|u_{\bar{T}} \odot v_{\bar{T}} - \theta^\star\| \leq \|\theta^\star_{:k} - S_{\bar{T}}\| + \|E_{\bar{T}}\| \lesssim \sqrt{d^2m}\alpha^{1-\mathcal{O}(\delta)} \vee \eta\theta^\star_1. \tag{67}$$

Therefore, with probability at least $1 - e^{-C\log^2(m)\log(d)\log(\|\theta^\star\|/\alpha)}$, we have

$$\|u_{\bar{T}} \odot v_{\bar{T}} - \theta^\star\| \lesssim \sqrt{d^2m}\alpha^{1-\tilde{\Theta}\left(\frac{k^2}{\sqrt{m(1-p)^2}}\right)} \vee \eta\theta^\star_1. \tag{68}$$

Long Escape Time

Based on the above analysis, it can be seen that, after $\bar{T}$ iterations, the residual term is the dominant term, which may cause the algorithm to diverge, as captured by Figure 1. We now show that the residual term will not diverge within $T' = \sqrt{\frac{m(1-p)^2}{k}}\bar{T}$. To this goal, recall that for $\forall k+1 \leq i \leq d$, we have

$$u^2_{i,t+1} + v^2_{i,t+1} \leq (1+3\eta\delta)\left(u^2_{i,t} + v^2_{i,t}\right). \tag{69}$$

Hence, for $t \leq \frac{1}{6\eta\delta}\log\left(\frac{1}{\alpha}\right)$, we have

$$u^2_{i,t} + v^2_{i,t} \lesssim \alpha\left(1+3\eta\delta\right)^{\frac{1}{6\eta\delta}\log\left(\frac{1}{\alpha}\right)} \leq \alpha e^{0.5\log\left(\frac{1}{\alpha}\right)} = \sqrt{\alpha}. \tag{70}$$

The proof is completed by noticing that $m = \tilde{\Omega}\left(\frac{k}{(1-p)^2\delta^2}\right)$.

Balanced Property

To prove the balanced property, we directly calculate the dynamic of the difference $u_{i,t} - v_{i,t}$. One can write

$$u_{i,t+1} - v_{i,t+1} = (u_{i,t} - v_{i,t})\left(1 - \eta\frac{\theta^\star_i - u_{i,t}v_{i,t}}{\|u_t \odot v_t - \theta^\star\|} - \eta\delta_i\right). \tag{71}$$

Since $\theta^\star_i - u_{i,t}v_{i,t} \geq 0$, we have

$$|u_{i,t+1} - v_{i,t+1}| \leq |u_{i,t} - v_{i,t}|\left(1+\eta\delta\right), \tag{72}$$

for $0 \leq i \leq k$. We conclude that

$$|u_{i,t} - v_{i,t}| \leq \sqrt{\alpha}\left(1+\eta\delta\right)^t \lesssim \alpha^{0.5-\Theta(k^{3/2}\delta)}. \tag{73}$$

for $\forall t \lesssim \frac{k^{3/2}}{\eta}\log\left(\frac{1}{\alpha}\right)$. On the other hand, for $i \geq k$, we can write

$$|u_{i,t} - v_{i,t}| \leq \sqrt{u^2_{i,t} + v^2_{i,t}} \lesssim \alpha^{0.5-\Theta(k^{3/2}\delta)}. \tag{74}$$

The proof is completed by noticing that $m = \tilde{\Omega}\left(\frac{k}{(1-p)^2\delta^2}\right)$.

Convergence in Under-parameterized Regime

In this section, we study the under-parameterized regime, where we assume that $m = \tilde{\Omega}(\frac{d}{(1-p)^2})$. The analysis of the signal term is the same as the over-parameterized regime and hence omitted for brevity. Here we only analyze the residual dynamic. One can write

$$E_{t+1} = E_t \left(\mathbf{1} - \eta \frac{u_{k+1:,t}^2 + v_{k+1:,t}^2}{\|u_t \odot v_t - \theta^\star\|} \right) + \eta \delta_{k+1:} \left(u_{k+1:,t}^2 + v_{k+1:,t}^2 \right) + \eta^2 \left(\frac{-E_t}{\|u_t \odot v_t - \theta^\star\|} + \delta_{k+1:} \right)^2 E_t. \tag{75}$$

When the residual term becomes the dominant term, i.e., $\|\theta_{:k}^\star - S_t\| \le \|E_t\|$, we have the simplified dynamic

$$\begin{aligned} \|E_{t+1}\| &\le \left\| E_t \left(\mathbf{1} - 0.5\eta \frac{u_{k+1:,t}^2 + v_{k+1:,t}^2}{\|u_t \odot v_t - \theta^\star\|} \right) + \eta \delta_{k+1:} \left(u_{k+1:,t}^2 + v_{k+1:,t}^2 \right) \right\| + 2\eta^2 \left(\|E_t\|^3 + \delta^2 \|E_t\| \right) \\ &\overset{(a)}{\le} \left\| E_t \left(\mathbf{1} - \frac{\eta E_t}{\|E_t\|} \right) \right\| + 4\eta\delta \|E_t\| \\ &\le \left(1 - \frac{\eta}{\sqrt{d}} \right) \|E_t\| + 4\eta\delta \|E_t\| \\ &\overset{(b)}{\le} \left(1 - 0.5\frac{\eta}{\sqrt{d}} \right) \|E_t\|. \end{aligned} \tag{76}$$

Here in (a) we used the balanced property, which results in $u_{k+1:,t}^2 + v_{k+1:,t}^2 \asymp 2E_t$. Moreover, (b) is implied by the fact that $m \gtrsim \frac{d}{(1-p)^2}$, which in turn implies $\delta \lesssim \frac{1}{\sqrt{d}}$. Hence, we have

$$\|u_{t+1} \odot v_{t+1} - \theta^\star\| \le \left(1 - \Omega\left(\frac{\eta}{\sqrt{d}} \right) \right) \|u_t \odot v_t - \theta^\star\|. \tag{77}$$

Then, for $t \ge \bar{T}$, we have

$$\|u_t \odot v_t - \theta^\star\| \lesssim \sqrt{d^2 m \alpha}^{\,1-\tilde{\Theta}\left(\frac{k^2}{\sqrt{m(1-p)^2}}\right)} \left(1 - \Omega\left(\frac{\eta}{\sqrt{d}} \right) \right)^{t-\bar{T}} \vee \eta\theta_1^\star. \tag{78}$$

C.2 Proof of Theorem 3

The proof of N-layer model is similar to that of 2-layer model. First, we study the signal dynamics, showing that the first k components $\prod w_{i,j}^{(t)}$ converge to $\theta_j^\star$ sequentially for $1 \le j \le k$. We also prove that the residual term remains small along the optimization trajectory. Based on the SubGM update rule, one can write

$$w_i^{(t+1)} = w_i^{(t)} + \eta \frac{\theta^\star - \prod w_j^{(t)}}{\left\| \theta^\star - \prod w_j^{(t)} \right\|} \prod_{j \ne i} w_j^{(t)} + \eta \boldsymbol{\delta} \prod_{j \ne i} w_j^{(t)}, \text{ and } \|\boldsymbol{\delta}\|_\infty \le \delta, \forall i \in [N]. \tag{79}$$

Signal Dynamics

Stage 1: In this stage, we show that $\prod w_{i,1}$ will converge to $\theta^\star$ within $T_1 = \Theta\left(\frac{\|\theta^\star\|}{N\eta\theta_1^\star}\alpha^{-\frac{N-2}{N}}\right)$ iterations. We first prove the upper bound. According to the update rule, we have

$$\begin{aligned}
\prod w_{i,1}^{(t+1)} &= \prod w_{i,1}^{(t)} + \sum_{i=1}^{N} \eta \left(\frac{\theta_1^\star - \prod w_{j,1}^{(t)}}{\left\| \theta^\star - \prod w_j^{(t)} \right\|} + \delta_1 \right) \left(\prod_{j\neq i} w_{j,1}^{(t)} \right)^2 + \text{higher order terms of } \eta \\
&\geq \prod w_{i,1}^{(t)} + \sum_{i=1}^{N} \eta \left(\frac{\theta_1^\star - \prod w_{j,1}^{(t)}}{\left\| \theta^\star - \prod w_j^{(t)} \right\|} + \delta_1 \right) \left(\prod_{j\neq i} w_{j,1}^{(t)} \right)^2 \qquad (80) \\
&\geq \prod w_{i,1}^{(t)} + N\eta \left(\frac{\theta_1^\star - \prod w_{j,1}^{(t)}}{\left\| \theta^\star - \prod w_j^{(t)} \right\|} - \delta \right) \left(\prod w_{i,1}^{(t)} \right)^{\frac{2(N-1)}{N}}.
\end{aligned}$$

Here we use the fact that $\prod w_{i,1}^{(t)} \leq \theta_1^\star$ and $\eta \lesssim \frac{1}{N}\frac{1}{\kappa}^{\frac{N-2}{N}}$ so that we can drop the higher order terms. For brevity, we only show how we can drop the 2-th order term of η. The proof of the higher order terms is similar. One can write

$$\begin{aligned}
\eta^2 \sum_{i\neq j} \left(\prod_{k\neq i} w_{k,1}^{(t)} \prod_{k\neq j} w_{k,1}^{(t)} \prod_{k\neq i,j} w_{k,1}^{(t)} \right) &\overset{(a)}{\leq} \eta^2 \left(\theta_1^\star\right)^{\frac{N-2}{N}} \sum_{i\neq j} \left(\prod_{k\neq i} w_{k,1}^{(t)} \prod_{k\neq j} w_{k,1}^{(t)} \right) \\
&\overset{(b)}{\leq} (N-1)\eta \left(\theta_1^\star\right)^{\frac{N-2}{N}} \eta \sum_{i=1}^{m} \left(\prod_{j\neq i} w_{j,1}^{(t)} \right)^2 \qquad (81) \\
&\overset{(c)}{\lesssim} \eta \sum_{i=1}^{m} \left(\prod_{j\neq i} w_{j,1}^{(t)} \right)^2.
\end{aligned}$$

Here in (a) we use the balanced property and the fact that $\prod w_{i,1}^{(t)} \leq \theta_1^\star$. Moreover, in (b), we use the rearrangement Inequality. Finally in (c) we use the assumption that $\eta \lesssim N^{-1}\kappa^{-\frac{N-2}{N}}$. For simplicity, we denote $x_t = \prod w_{i,1}^{(t)}$. Note that $\left\| \theta^\star - \prod w_j^{(t)} \right\| \leq \|\theta^\star\|$. Hence, the dynamic can be simplified as

$$x_{t+1} \geq x_t + \frac{N\eta}{\|\theta^\star\|} \left(\theta_1^\star - \delta \|\theta^\star\| - x_t \right) x_t^{\frac{2(N-1)}{N}}. \qquad (82)$$

We next show that $x_T \geq \theta_1^\star - 2\delta\|\theta^\star\|$ within $T_1 = \Theta\left(\frac{\|\theta^\star\|}{N\eta\theta_1^\star}\alpha^{-\frac{N-2}{N}}\right)$ iterations provided that $x_0 = \Theta(\alpha)$. To this goal, we divide our analysis into two substages.

- $x_t \leq \frac{\theta_1^\star}{2}$: In this substage, we assume that $x_t \leq \frac{\theta_1^\star}{2}$. Hence, we can further simplify the dynamic as

$$x_{t+1} \geq x_t + 0.5N\eta \frac{\theta_1^\star}{\|\theta^\star\|} x_t^{\frac{2(N-1)}{N}}. \qquad (83)$$

Without loss of generality, we assume that $x_0 = \alpha$. Now we divide the interval $[\alpha, 0.5\theta_1^\star]$ into a series of sub-intervals $\{\mathcal{I}_k\}$, where $\mathcal{I}_k = [2^k\alpha, 2^{k+1}\alpha)$. In each $\mathcal{I}_k$, the dynamic can be further simplified as

$$x_{t+1} \geq \left(1 + 0.5N\eta \frac{\theta_1^\star}{\|\theta^\star\|} \left(2^k \alpha\right)^{\frac{N-2}{N}}\right) x_t. \tag{84}$$

Therefore, the number of iterations that x_t spends in each interval $\mathcal{I}_k$ is $\mathcal{O}\left(\frac{\|\theta^\star\|}{N\eta\theta_1^\star}\left(2^k\alpha\right)^{-\frac{N-2}{N}}\right)$. Hence, the total number of iterations is upper bounded by $\mathcal{O}\left(\sum_{k=0}^{\infty} \frac{\|\theta^\star\|}{N\eta\theta_1^\star}\left(2^k\alpha\right)^{-\frac{N-2}{N}}\right) = \mathcal{O}\left(\frac{\|\theta^\star\|}{N\eta\theta_1^\star}\alpha^{-\frac{N-2}{N}}\right)$.

- $x_t \geq \frac{\theta_1^\star}{2}$: In this substage, we define $y_t = \theta_1^\star - \delta\|\theta^\star\| - x_t$. Via a similar trick, we can show that within additional $\mathcal{O}\left(\frac{\|\theta^\star\|}{N\eta}(\theta_1^\star)^{-\frac{2N-2}{N}}\right)$ iterations, we have $x_t \geq \theta_1^\star - 2\delta\|\theta^\star\|$. Overall, after $T_1 = \Theta\left(\frac{\|\theta^\star\|}{N\eta\theta_1^\star}\alpha^{-\frac{N-2}{N}}\right)$ iterations, we have $\theta_1^\star - 2\delta\|\theta^\star\| \leq \prod w_{i,1}^{(T_1)} \leq \theta_1^\star$.

Stages 2 to k: Similarly, for component $\prod w_{j,i}^{(t)}$, it takes $\mathcal{O}\left(\frac{\left\|\theta_{-(i-1)}^\star\right\|}{N\eta\theta_i^\star}\alpha^{-\frac{N-2}{N}}\right)$ iterations to attain $\theta_i^\star - 2\delta\|\theta^\star\|$. Overall, Stages 2 to k take $\Theta\left(\frac{k^{\frac{3}{2}}}{N\eta}\alpha^{-\frac{N-2}{N}}\right)$ iterations to terminate.

Stage $k+1$: In this stage, we take $S_t = \prod w_{j,:k}^{(t)}$. Hence, we have

$$\begin{aligned}
\|\theta_{:k}^\star - S_{t+1}\| &\leq \left\|(\theta_{:k}^\star - S_t)\left(\mathbf{1} - \eta\frac{\sum_{i=1}^N\left(\prod_{j\neq i} w_{j,1}^{(t)}\right)^2}{\left\|\theta^\star - \prod w_j^{(t)}\right\|}\right)\right\| + 4N\eta\delta\sqrt{k}(\theta_1^\star)^{\frac{2(N-1)}{N}} \\
&\leq \|\theta_{:k}^\star - S_t\|\left(1 - N\eta\frac{(\theta_k^\star)^{\frac{2(N-1)}{N}}}{\left\|\theta_{:k}^\star - S_t\right\| + \|E_t\|}\right) + 4N\eta\delta\sqrt{k}(\theta_1^\star)^{\frac{2(N-1)}{N}} \\
&\leq \|\theta_{:k}^\star - S_t\| - 0.5N\eta(\theta_k^\star)^{\frac{2(N-1)}{N}} + 4N\eta\delta\sqrt{k}(\theta_1^\star)^{\frac{2(N-1)}{N}} \\
&\leq \|\theta_{:k}^\star - S_t\| - 0.1N\eta(\theta_k^\star)^{\frac{2(N-1)}{N}}.
\end{aligned} \tag{85}$$

Here we used the fact that $\left\|\theta^\star - \prod w_j^{(t)}\right\| \leq \|\theta_{:k}^\star - S_t\| + \|E_t\| \leq 2\|\theta_{:k}^\star - S_t\|$, and the assumption that $\delta \lesssim \frac{1}{N}\frac{1}{\kappa}^{\frac{2N-2}{N}}$. On the other hand, we have

$$\begin{aligned}
\|S_{t+1} - S_t\| &\leq \left\|\sum_{i=1}^N \eta\left(\frac{\theta_{:k}^\star - S_t}{\left\|\theta^\star - \prod w_j^{(t)}\right\|} + \boldsymbol{\delta}_{i,:k}\right)\left(\prod_{j\neq i} w_{j,:k}^{(t)}\right)^2\right\| \\
&\leq 2N\eta\sqrt{k}(\theta_1^\star)^{\frac{2(N-1)}{N}}.
\end{aligned} \tag{86}$$

Hence, we conclude that within $\mathcal{O}\left(\frac{\sqrt{k}\delta}{N\eta}\frac{1}{\kappa}^{\frac{2(N-1)}{N}}\right)$ iterations, we have $\|\theta_{:k}^\star - S_t\| \lesssim \sqrt{d^2m}\alpha \vee N\eta(\theta_1^\star)^{\frac{2(N-1)}{N}}$. Overall, the total iteration complexity is bounded by $\mathcal{O}\left(\frac{k^{\frac{3}{2}}}{N\eta}\alpha^{-\frac{N-2}{N}}\right)$.

Residual Dynamics

Similar to the 2-layer model, here we study the surrogate of the residual term $\sum_{i=1}^N \left(w_{i,l}^{(t)}\right)^2$ for $l \geq k+1$. To this goal, we first notice that

$$\begin{aligned}
\sum_{i=1}^N \left(w_{i,l}^{(t+1)}\right)^2 = \sum_{i=1}^N \left(w_{i,l}^{(t)}\right)^2 &+ 2N\eta \left(\frac{-\prod w_{j,l}^{(t)}}{\left\|\theta^\star - \prod w_j^{(t)}\right\|} + \delta_{i,l}\right) \prod w_{j,l}^{(t)} \\
&+ \eta^2 \sum_{i=1}^N \left(\frac{-\prod w_{j,l}^{(t)}}{\left\|\theta^\star - \prod w_j^{(t)}\right\|} + \delta_{i,l}\right)^2 \left(\prod_{j\neq i} w_{j,l}^{(t)}\right)^2 \\
&\leq \sum_{i=1}^N \left(w_{i,l}^{(t)}\right)^2 + 4N\eta\delta \prod w_{j,l}^{(t)} \\
&\leq \sum_{i=1}^N \left(w_{i,l}^{(t)}\right)^2 + 4N\eta\delta \left(\frac{\sum_{i=1}^N \left(w_{i,l}^{(t)}\right)^2}{N}\right)^{\frac{N}{2}}.
\end{aligned} \tag{87}$$

Hence, once we set $z_t = \sum_{i=1}^N \left(w_{i,l}^{(t)}\right)^2$, we have the following simplified dynamic

$$z_{t+1} \leq z_t + 4N\eta\delta \left(\frac{z_t}{N}\right)^{\frac{N}{2}}, \tag{88}$$

with $z_0 = \Theta\left(N\alpha^{\frac{2}{N}}\right)$. We claim that within $\mathcal{O}\left(\frac{1}{N\eta\alpha}\right)$ iterations, we still have $z_t = \Theta\left(N\alpha^{\frac{2}{N}}\right)$. To show this, we suppose without loss of generality that $z_0 = N\alpha^{\frac{2}{N}}$, and define T as the first time that $z_T \geq 2N\alpha^{\frac{2}{N}}$. For any $0 \leq t \leq T-1$, we have

$$z_{t+1} \leq z_t + 4N\eta\delta 2^{\frac{N}{2}}\alpha. \tag{89}$$

We conclude that

$$T \geq \frac{N\alpha^{\frac{2}{N}}}{4N\eta\delta 2^{\frac{N}{2}}\alpha} = \frac{1}{4\delta 2^{\frac{N}{2}}} \frac{1}{\eta} \alpha^{-\frac{N-2}{N}} \gtrsim \frac{1}{N\eta}\alpha^{-\frac{N-2}{N}}. \tag{90}$$

Therefore, via a basic inequality, we have

$$\prod w_{i,l}^{(t)} \leq \left(\frac{\sum_{i=1}^N \left(w_{i,l}^{(t)}\right)^2}{N}\right)^{\frac{N}{2}} \lesssim \alpha. \tag{91}$$

Combining the analysis of both signal and residual terms, we conclude that within $\Theta\left(\frac{1}{N\eta}\alpha^{-\frac{N-2}{N}}\right)$ iterations, we have

$$\left\|\prod w_i^{(t)} - \theta^\star\right\| \lesssim \sqrt{d^2 m \alpha} \vee \left(\eta\theta_1^\star\right)^{\frac{2(N-1)}{N}}. \tag{92}$$

Long Time Guarantee

Similar to the proof of the 2-layer model, one can show that the residual term becomes the dominant term in the generalization error, and it stays in the order of α within $\Omega\left(\frac{1}{N\eta\delta}\alpha^{-\frac{N-2}{N}}\right)$ iterations. The details are omitted for brevity.

Balanced Property

To prove the balanced property, we first study the dynamic of $w_{i,l}^{(t)} - w_{j,l}^{(t)}$, $\forall l \in [k], i, j \in [N]$. To this goal, we have

$$w_{i,l}^{(t+1)} - w_{j,l}^{(t+1)} = \left(w_{i,l}^{(t)} - w_{j,l}^{(t)}\right)\left(1 - \eta\left(\frac{\theta_l^\star - \prod w_{j,l}^{(t)}}{\left\|\theta^\star - \prod w_j^{(t)}\right\|} + \delta_l\right)\prod_{f\neq i,j} w_{f,l}^{(t)}\right), \tag{93}$$

which in turn implies

$$\left|w_{i,l}^{(t+1)} - w_{j,l}^{(t+1)}\right| \leq \left|w_{i,l}^{(t)} - w_{j,l}^{(t)}\right|\left(1 - \eta\left(\frac{\theta^\star - \prod w_j^{(t)}}{\left\|\theta^\star - \prod w_j^{(t)}\right\|} + \delta_l\right)\prod_{f\neq i,j} w_f^{(t)}\right). \tag{94}$$

If $\prod w_{j,l}^{(t)} \leq \theta_l^\star - \delta\left\|\theta^\star\right\|$, the above inequality can be simplified as

$$\left|w_{i,l}^{(t)} - w_{j,l}^{(t)}\right| \leq \left|w_{i,l}^{(0)} - w_{j,l}^{(0)}\right| \lesssim \alpha^{\frac{1}{N}}, \forall i, j \in [N], l \in [k]. \tag{95}$$

Once $\prod w_{j,l}^{(t)} \geq \theta_l^\star - \delta\left\|\theta^\star\right\|$, we immediately have $w_{j,l}^{(t)} = \sqrt[N]{\theta_l^\star} \pm \mathcal{O}(\sqrt[N]{\alpha})$. Then, we show that $w_{j,l}^{(t)}$ will stay close to $\sqrt[N]{\theta_l^\star}$. To this goal, we first observe that $\left(w_{i,l}^{(t+1)} - w_{i,l}^{(t)}\right)w_{i,l}^{(t)} \equiv \left(w_{j,l}^{(t+1)} - w_{j,l}^{(t)}\right)w_{j,l}^{(t)}$, $\forall i, j \in [N]$, which indicates that $w_{i,l}^{(t)}$ increases or decreases simultaneously. Hence, we conclude that $\left|w_{i,l}^{(t)} - w_{j,l}^{(t)}\right| \lesssim \delta\sqrt[N]{\theta_l^\star}$.

For the residual term, we can derive a tighter bound. First, we have

$$\left(w_{i,l}^{(t+1)}\right)^2 = \left(w_{i,l}^{(t)}\right)^2 + 2\eta\left(\frac{-\prod w_{j,l}^{(t)}}{\left\|\theta^\star - \prod w_j^{(t)}\right\|} + \delta_l\right)\prod w_{j,l}^{(t)} + \eta^2\left(\frac{-\prod w_{j,l}^{(t)}}{\left\|\theta^\star - \prod w_j^{(t)}\right\|} + \delta_l\right)^2\left(\prod_{j\neq i} w_{j,l}^{(t)}\right)^2. \tag{96}$$

Since we have already shown that $\prod w_{i,l}^{(t)} \lesssim \alpha$, we further have

$$\left(w_{i,l}^{(t+1)}\right)^2 \leq \left(w_{i,l}^{(t)}\right)^2 + 4\eta\delta\alpha. \tag{97}$$

Therefore, one can write $\left(w_{i,l}^{(t)}\right)^2 \lesssim \left(w_{i,l}^{(0)}\right)^2 + 4\eta\delta\alpha\frac{1}{\eta}\alpha^{-\frac{N-2}{N}} \lesssim \alpha^{\frac{2}{N}}$, which in turn implies $\left|w_{i,l}^{(t)} - w_{j,l}^{(t)}\right| \leq \left|w_{i,l}^{(t)}\right| + \left|w_{j,l}^{(t)}\right| \lesssim \alpha^{1/N}$.

Convergence in Under-parameterized Regime

Similar to the 2-layer model, we consider the dynamic of $E_t = \prod w_{i,k+1:}^{(t)}$, which is characterized as follows

$$\begin{aligned}
\|E_{t+1}\| &\le \left\| E_t + \sum_{i=1}^{N} \eta \left(\frac{-E_t}{\left\|\theta^\star - \prod w_j^{(t)}\right\|} + \delta_{k+1:} \right) \left(\prod_{j\neq i} w_{j,k+1:}^{(t)} \right)^2 \right\| \\
&\le \left\| E_t + N\eta \left(\frac{-E_t}{\left\|\theta^\star - \prod w_j^{(t)}\right\|} + \delta \right) E_t^{\frac{2(N-1)}{N}} \right\| \\
&\le \left\| E_t + N\eta \left(\frac{-E_t}{\|E_t\|} + \delta \right) E_t^{\frac{2(N-1)}{N}} \right\| \\
&\le \|E_t\| - N\eta d^{-\frac{N-1}{N}} \|E_t\|^{\frac{2N-2}{N}} + N\eta\delta \left\| E_t^{\frac{2(N-1)}{N}} \right\| \\
&\le \|E_t\| - N\eta d^{-\frac{N-1}{N}} \|E_t\|^{\frac{2N-2}{N}} + N\eta\delta \|E_t\|^{\frac{2N-2}{N}} \\
&\le \|E_t\| - N\eta d^{-\frac{N-1}{N}} \|E_t\|^{\frac{2N-2}{N}} .
\end{aligned} \tag{98}$$

The last inequality comes from the fact that $\delta \lesssim d^{-\frac{N-1}{N}}$ since we assume $m \gtrsim \frac{d^{\frac{2N-2}{N}}}{(1-p)^2}$. Hence, we have

$$\|E_t\| \lesssim \left(\frac{1}{N\eta d^{-(N-1)/N}(t-\bar{T}) + 1/\|E_{\bar{T}}\|} \right)^{\frac{N}{N-2}} . \tag{99}$$

Since the residual term is the dominant term in the generalization error, we have

$$\left\|\prod w_i^{(t)} - \theta^\star\right\| \lesssim \left(\frac{\left\|\prod w_i^{(\bar{T})} - \theta^\star\right\|}{\left\|\prod w_i^{(\bar{T})} - \theta^\star\right\| N\eta d^{-(N-1)/N}(t-\bar{T}) + 1} \right)^{\frac{N}{N-2}} , \tag{100}$$

which completes the proof.

D Proof of Proposition 1

First, we provide an upper bound of the covering number for the (k,ϑ)-approximate sparse unit ball. We defer a preliminary discussion on covering number to Appendix G .

Lemma 4. *Let $\mathcal{T}_{k,\vartheta} := \{u \in \mathbf{R}^d : u \text{ is } (k,\vartheta)\text{-approximate sparse}, \|u\| \le 1\}$. Then its covering number $N(\mathcal{T}_{k,\vartheta}, \epsilon, \|\cdot\|)$ is upper bounded by*

$$N(\mathcal{T}_{k,\vartheta}, \epsilon, \|\cdot\|) \le \left(\frac{ed}{k}\right)^k \left(1 + \frac{4}{\epsilon}\right)^k , \tag{101}$$

provided that $\epsilon \ge \vartheta$.

The next lemma will play a crucial role in proving Proposition 1.

Lemma 5. *Suppose $x \in \mathbf{R}^d$ is a standard Gaussian vector, i.e., $x_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, and the noise ϵ satisfies Assumption 1, then we have*

$$\varphi(u) = \frac{\mathbb{E}\left[\mathrm{Sign}\left(\langle x, u\rangle + \epsilon\right)\langle x, v\rangle\right]}{\left\langle \frac{u}{\|u\|}, v\right\rangle} = \sqrt{\frac{2}{\pi}}(1-p) + \sqrt{\frac{2}{\pi}}p\mathbb{E}\left[e^{-\epsilon^2/(2\|u\|^2)}\right].$$

The proof of this lemma can be found in Appendix F.1. Now, we are ready to prove Proposition 1. Our goal is to show that for arbitrary $u \in \mathcal{A}$, the following inequality holds

$$\left\|\frac{1}{m}\sum_{i=1}^{m}\mathrm{Sign}\left(\langle x_i, u\rangle + \epsilon_i\right)x_i - \varphi(u)\frac{u}{\|u\|}\right\|_\infty \le \delta \tag{102}$$

with probability at least $1 - Ce^{-cm\delta^2}$ provided that $m \gtrsim \frac{k\log(d)\log(R)\log(\frac{1}{\vartheta})}{(1-p)^2}$. Here we define $\mathcal{A} := \{u : r \le \|u\| \le R, u \text{ is } (k,\vartheta)\text{-approximate sparse}\}$, where $r \gtrsim \sqrt{dm/k}\vartheta\log(1/\vartheta)$. Moreover, we define $\mathcal{B} := \{u : r \le \|u\| \le R, \|u\|_0 \le k\}$ and $\mathcal{C} := \{(u, u') : u \in \mathcal{A}, v \in \mathcal{B}_\zeta, \|u - u'\| \le \zeta\}$. Here $\mathcal{B}_\zeta$ is the ζ-net of $\mathcal{B}$ with $\zeta \gtrsim r$. Finally, we define $\mathcal{D} := \{\pm\mathbf{e}_j\}_{j\in[d]}$, where $\mathbf{e}_j$ forms the standard basis of $\mathbf{R}^d$. Based on these definitions, we have

$$\begin{aligned}
&\sup_{u\in\mathcal{A}}\left\|\frac{1}{m}\sum_{i=1}^{m}\mathrm{Sign}\left(\langle x_i, u\rangle + \epsilon_i\right)x_i - \varphi(u)\frac{u}{\|u\|}\right\|_\infty \\
&= \sup_{u\in\mathcal{A}, v\in\mathcal{D}}\frac{1}{m}\sum_{i=1}^{m}\mathrm{Sign}\left(\langle x_i, u\rangle + \epsilon_i\right)\langle x_i, v\rangle - \frac{\varphi(u)}{\|u\|}\langle u, v\rangle \\
&= \sup_{v\in\mathcal{D}}\left\{\sup_{u\in\mathcal{A}}\frac{1}{m}\sum_{i=1}^{m}\mathrm{Sign}\left(\langle x_i, u\rangle + \epsilon_i\right)\langle x_i, v\rangle - \frac{\varphi(u)}{\|u\|}\langle u, v\rangle\right\}.
\end{aligned} \tag{103}$$

We then show that for each element $y \in \mathcal{D}$, $\sup_{u\in\mathcal{A}}\frac{1}{m}\sum_{i=1}^{m}\mathrm{Sign}\left(\langle x_i, u\rangle + \epsilon_i\right)\langle x_i, y\rangle - \varphi(u)\frac{\langle u, y\rangle}{\|u\|}$, $j \in [d]$ is $\mathcal{O}\left(\frac{1}{m}\right)$-sub-Gaussian random variable. To see this, note that

$$\begin{aligned}
\left\|\mathrm{Sign}\left(\langle x_i, u\rangle + \epsilon_i\right)x_{i,j} - \varphi(u)\frac{u_j}{\|u\|}\right\|_{\psi_2} &\le \left\|\mathrm{Sign}\left(\langle x_i, u\rangle + \epsilon_i\right)x_{i,j}\right\|_{\psi_2} + \sqrt{\frac{2}{\pi}} \\
&\le \|x_{i,j}\|_{\psi_2} + \sqrt{\frac{2}{\pi}} = \mathcal{O}(1).
\end{aligned} \tag{104}$$

Here we use the property of sub-Gaussian norm. This implies that $\frac{1}{m}\sum_{i=1}^{m}\mathrm{Sign}\left(\langle x_i, u\rangle + \epsilon_i\right)x_{i,j} - \varphi(u)\frac{u_j}{\|u\|}$ is $\mathcal{O}\left(\frac{1}{m}\right)$-sub-Gaussian random variable, since it is the sample average of $\mathrm{Sign}\left(\langle x_i, u\rangle + \epsilon_i\right)x_{i,j} - \varphi(u)\frac{u_j}{\|u\|}$.

Hence, via maximal inequality, we have that for $\forall t > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{u\in\mathcal{A}}\left\|\frac{1}{m}\sum_{i=1}^{m}\mathrm{Sign}\left(\langle x_i, u\rangle + \epsilon_i\right)x_i - \varphi(u)\frac{u}{\|u\|}\right\|_\infty\right. \\
&\qquad\left.\ge \sup_{y\in\mathcal{D}}\mathbb{E}\left[\sup_{u\in\mathcal{A}}\frac{1}{m}\sum_{i=1}^{m}\mathrm{Sign}\left(\langle x_i, u\rangle + \epsilon_i\right)\langle x_i, y\rangle - \varphi(u)\frac{\langle u, y\rangle}{\|u\|}\right] + t\right) \\
&\le 2de^{-cmt^2}.
\end{aligned} \tag{105}$$

Hence, it suffices to study $\mathbb{E}\left[\sup_{u\in\mathcal{A}}\frac{1}{m}\sum_{i=1}^{m}\mathrm{Sign}\left(\langle x_i,u\rangle+\epsilon_i\right)x_{i,1}-\varphi(u)\frac{u_1}{\|u\|}\right]$. To this goal, we decompose it into two terms via triangle inequality.

$$\mathbb{E}\left[\sup_{u\in\mathcal{A}}\frac{1}{m}\sum_{i=1}^{m}\mathrm{Sign}\left(\langle x_i,u\rangle+\epsilon_i\right)x_{i,1}-\varphi(u)\frac{u_1}{\|u\|}\right]\le(\mathrm{A})+(\mathrm{B}),\tag{106}$$

where

$$(\mathrm{A}):=\mathbb{E}\left[\sup_{u\in\mathcal{B}_\zeta}\frac{1}{m}\sum_{i=1}^{m}\mathrm{Sign}\left(\langle x_i,u\rangle+\epsilon_i\right)x_{i,1}-\varphi(u)\frac{u_1}{\|u\|}\right],\tag{107}$$

and

$$(\mathrm{B}):=\mathbb{E}\left[\sup_{(u,u')\in\mathcal{C}}\frac{1}{m}\sum_{i=1}^{m}\left(\mathrm{Sign}\left(\langle x_i,u\rangle+\epsilon_i\right)-\mathrm{Sign}\left(\left\langle x_i,u'\right\rangle+\epsilon_i\right)\right)x_{i,1}-\varphi(u)\frac{u_1}{\|u\|}+\varphi(u')\frac{u_1'}{\|u'\|}\right].\tag{108}$$

We first control (A). To this goal, we apply the union bound. Note that $\frac{1}{m}\sum_{i=1}^{m}\mathrm{Sign}\left(\langle x_i,u\rangle+\epsilon_i\right)x_{i,1}-\varphi(u)\frac{u_1}{\|u\|}$ is $\mathcal{O}(\frac{1}{m})$-sub-Gaussian and $|\mathcal{B}_\zeta|\le\left(\frac{R}{\zeta}\right)^{Ck\log(d)}$. We then have

$$(\mathrm{A})\lesssim\sqrt{\frac{k\log(d)\log\left(\frac{R}{\zeta}\right)}{m}}.\tag{109}$$

Now we control (B). Via triangle inequality, we first obtain

$$\begin{aligned}(\mathrm{B})\le&\underbrace{\mathbb{E}\left[\sup_{(u,u')\in\mathcal{C}}\frac{1}{m}\sum_{i=1}^{m}\left(\mathrm{Sign}\left(\langle x_i,u\rangle+\epsilon_i\right)-\mathrm{Sign}\left(\left\langle x_i,u'\right\rangle+\epsilon_i\right)\right)x_{i,1}\right]}_{(\mathrm{B}_1)}\\&+\underbrace{\sup_{(u,u')\in\mathcal{C}}\left\{-\varphi(u)\frac{u_1}{\|u\|}+\varphi(u')\frac{u_1'}{\|u'\|}\right\}}_{(\mathrm{B}_2)}.\end{aligned}\tag{110}$$

For the first part, applying Hölder's inequality leads to

$$\begin{aligned}(\mathrm{B}_1)\le&\,\mathbb{E}\left[\sup_{(u,u')\in\mathcal{C}}\left(\frac{1}{m}\sum_{i=1}^{m}\left|\mathrm{Sign}\left(\langle x_i,u\rangle+\epsilon_i\right)-\mathrm{Sign}\left(\left\langle x_i,u'\right\rangle+\epsilon_i\right)\right|\right)\max_{1\le i\le m}|x_{i,1}|\right]\\\le&\,\mathbb{E}\left[\sup_{(u,u')\in\mathcal{C}}\left(\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}\left(\left|\left\langle x_i,u-u'\right\rangle\right|\ge\left|\langle x_i,u\rangle+\epsilon_i\right|\right)\right)\max_{1\le i\le m}|x_{i,1}|\right]\\\le&\underbrace{\mathbb{E}\left[\sup_{\|\Delta u\|\le\zeta}\left(\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}\left(\left|\langle x_i,\Delta u\rangle\right|\ge t\right)\right)\max_{1\le i\le m}|x_{i,1}|\right]}_{(\mathrm{B}_3)}\\&+\underbrace{\mathbb{E}\left[\sup_{u\in\mathcal{B}_\zeta}\left(\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}\left(\left|\langle x_i,u\rangle+\epsilon_i\right|\le t\right)\right)\max_{1\le i\le m}|x_{i,1}|\right]}_{(\mathrm{B}_4)},\end{aligned}\tag{111}$$

where $t > 0$ is a constant to be determined later. Here, we used the fact that $\mathbb{1}\left(\left|\left\langle x_i, u - u'\right\rangle\right| \geq \left|\left\langle x_i, u\right\rangle + \epsilon_i\right|\right) \leq \mathbb{1}\left(\left|\left\langle x_i, \Delta u\right\rangle\right| \geq t\right) + \mathbb{1}\left(\left|\left\langle x_i, u\right\rangle + \epsilon_i\right| \leq t\right)$ in the last inequality. We first bound (B$_3$)

$$\begin{aligned}
(\mathrm{B}_3) &\leq \mathbb{E}\left[\left(\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}\left(\zeta\left\|x_i\right\| \geq t\right)\right)\max_{1\leq i\leq m}\left|x_{i,1}\right|\right] \\
&\leq \mathbb{E}\left[\mathbb{1}\left(\zeta\left\|x_i\right\| \geq t\right)\right]\mathbb{E}\left[\max_{j\neq i}\left|x_{j,1}\right|\right] + \mathbb{E}\left[\mathbb{1}\left(\zeta\left\|x_i\right\| \geq t\right)\left|x_{i,1}\right|\right] \\
&\lesssim e^{-C\frac{t^2}{\zeta^2}}\sqrt{\log(m)} + \mathbb{E}\left[\mathbb{1}\left(\zeta\left\|x_i\right\| \geq t\right)\left|x_{i,1}\right|\right],
\end{aligned} \tag{112}$$

provided that $\frac{t}{\zeta} \gtrsim \sqrt{d}$. Applying Cauchy-Schwarz inequality, we have

$$\mathbb{E}\left[\mathbb{1}\left(\zeta\left\|x_i\right\| \geq t\right)\left|x_{i,1}\right|\right] \leq \sqrt{\mathbb{P}\left(\zeta\left\|x_i\right\| \geq t\right)}\sqrt{\mathbb{E}\left[x_{i,1}^2\right]} \leq e^{-C\frac{t^2}{\zeta^2}}. \tag{113}$$

Hence, we conclude that (B$_3$) $\lesssim e^{-C\frac{t^2}{\zeta^2}}\sqrt{\log(m)}$. Next we control (B$_4$). Note that $\max_i\left|x_{i,1}\right|$ is $\mathcal{O}\left(\log(m)\right)$-sub-Gaussian. Via union bound, we have

$$(\mathrm{B}_4) \leq \sup_{u\in\mathcal{B}}\mathbb{E}\left[\mathbb{1}\left(\left|\left\langle x_i, u\right\rangle + \epsilon_i\right| \leq t\right)\max_{1\leq i\leq m}\left|x_{i,1}\right|\right] + C\sqrt{\frac{k\log(m)\log(d)\log(\frac{R}{\zeta})}{m}}. \tag{114}$$

For the first part, applying the similar decomposition method, we have

$$\begin{aligned}
\mathbb{E}\left[\mathbb{1}\left(\left|\left\langle x_i, u\right\rangle + \epsilon_i\right| \leq t\right)\max_{1\leq i\leq m}\left|x_{i,1}\right|\right] &\leq \mathbb{E}\left[\mathbb{1}\left(\left|\left\langle x_i, u\right\rangle + \epsilon_i\right| \leq t\right)\max_{j\neq i}\left|x_{i,1}\right|\right] \\
&\quad + \mathbb{E}\left[\mathbb{1}\left(\left|\left\langle x_i, u\right\rangle + \epsilon_i\right| \leq t\right)\left|x_{i,1}\right|\right] \\
&\lesssim \sqrt{\log(m)}\frac{t}{r}.
\end{aligned} \tag{115}$$

Hence, we conclude that (B$_4$) $\lesssim \sqrt{\log(m)}\frac{t}{r} + \sqrt{\frac{k\log(m)\log(d)\log(\frac{R}{\zeta})}{m}}$. For (B$_2$), we first have

$$\begin{aligned}
\left|-\varphi(u)\frac{u_1}{\|u\|} + \varphi(u')\frac{u'_1}{\|u'\|}\right| &= \left|\varphi(u') - \varphi(u)\right|\frac{|u_1|}{\|u\|} + \varphi(u')\left|\frac{u'_1}{\|u'\|} - \frac{u_1}{\|u\|}\right| \\
&\lesssim \left|\varphi(u') - \varphi(u)\right| + \zeta.
\end{aligned} \tag{116}$$

For the first part, we use Mean Value Theorem to write

$$\left|\varphi(u') - \varphi(u)\right| \leq \left\|\nabla\varphi(v)\right\|\left\|u' - u\right\| \leq \left\|\nabla\varphi(v)\right\|\zeta, \tag{117}$$

where v is a point between u and u'. Note that $\nabla\varphi(v) = \sqrt{\frac{2}{\pi}}p\mathbb{E}\left[\frac{\epsilon^2 v}{\|v\|^4}e^{-\frac{\epsilon^2}{2\|v\|^2}}\right]$. Hence, we have

$$\sup_{\|v\|\geq r}\left\|\nabla\varphi(v)\right\| \lesssim \sup_{\|v\|\geq r}\mathbb{E}\left[\frac{\epsilon^2}{\|v\|^3}e^{-\frac{\epsilon^2}{2\|v\|^2}}\right] \leq \frac{1}{r}\sup_{\|v\|\geq r}\mathbb{E}\left[\frac{\epsilon^2}{\|v\|^2}e^{-\frac{\epsilon^2}{2\|v\|^2}}\right] \lesssim \frac{1}{r}. \tag{118}$$

Overall, we have $(\mathrm{B}_2) \lesssim \frac{\zeta}{r}$, which results in

$$\begin{aligned} &\mathbb{E}\left[\sup_{u\in\mathcal{A}} \frac{1}{m}\sum_{i=1}^m \mathrm{Sign}\left(\langle x_i, u\rangle + \epsilon_i\right) x_{i,1} - \varphi(u)\frac{u_1}{\|u\|}\right] \\ &\lesssim \frac{\zeta}{r} + e^{-C\frac{t^2}{\zeta^2}}\sqrt{\log\left(m\right)} + \sqrt{\log(m)}\frac{t}{r} + \sqrt{\frac{k\log(m)\log(d)\log(\frac{R}{\zeta})}{m}}. \end{aligned} \tag{119}$$

Hence, once we set $\zeta \asymp \vartheta$, and $t \asymp \sqrt{d}\vartheta\log(m)$, together with the assumption that $r \gtrsim \sqrt{\frac{dm}{k}\vartheta\log\left(\frac{1}{\vartheta}\right)}$, we conclude that

$$\mathbb{E}\left[\sup_{u\in\mathcal{A}} \frac{1}{m}\sum_{i=1}^m \mathrm{Sign}\left(\langle x_i, u\rangle + \epsilon_i\right) x_{i,1} - \varphi(u)\frac{u_1}{\|u\|}\right] \lesssim \sqrt{\frac{k\log^2(m)\log(d)\log(\frac{R}{\vartheta})}{m}}. \tag{120}$$

This leads to

$$\begin{aligned} &\mathbb{P}\left(\sup_{u\in\mathcal{A}}\left\|\frac{1}{m}\sum_{i=1}^m \mathrm{Sign}\left(\langle x_i, u\rangle + \epsilon_i\right) x_i - \varphi(u)\frac{u}{\|u\|}\right\|_\infty \geq C\sqrt{\frac{k\log^2(m)\log(d)\log(\frac{R}{\vartheta})}{m}} + \delta\right) \\ &\leq 2de^{-cm\delta^2}. \end{aligned} \tag{121}$$

Therefore, the following inequality holds, provided that $m \gtrsim \frac{k\log^2(m)\log(d)\log(\frac{R}{\vartheta})}{(1-p)^2\delta^2}$

$$\mathbb{P}\left(\sup_{u\in\mathcal{A}}\left\|\frac{1}{\varphi(u)}\frac{1}{m}\sum_{i=1}^m \mathrm{Sign}\left(\langle x_i, u\rangle + \epsilon_i\right) x_i - \frac{u}{\|u\|}\right\|_\infty \geq \delta\right) \leq e^{-cm\delta^2}. \tag{122}$$

Now we turn to the case $m \gtrsim \frac{d}{(1-p)^2}$. Following the same technique, it suffices to bound $\mathbb{E}\left[\sup_{u\in\mathbf{R}^d}\frac{1}{m}\sum_{i=1}^m \mathrm{Sign}\left(\langle x_i, u\rangle + \epsilon_i\right) x_{i,1} - \varphi(u)\frac{u_1}{\|u\|}\right]$. To this goal, we first notice that

$$\begin{aligned} &\mathbb{E}\left[\sup_{u\in\mathbf{R}^d} \frac{1}{m}\sum_{i=1}^m \mathrm{Sign}\left(\langle x_i, u\rangle + \epsilon_i\right) x_{i,1} - \varphi(u)\frac{u_1}{\|u\|}\right] \\ &= \underbrace{\mathbb{E}\left[\sup_{\|u\|=1,\lambda\in\mathbf{R}} \frac{1}{m}\sum_{i=1}^m \mathrm{Sign}\left(\langle x_i, u\rangle + \lambda\epsilon_i\right) x_{i,1} - \varphi(u)u_1\right]}_{(A)}. \end{aligned} \tag{123}$$

Similarly, applying one-step discretization, we have

$$\begin{aligned} (A) \leq\ &\underbrace{\mathbb{E}\left[\sup_{u\in\mathbb{S}_\epsilon,\lambda\in\mathbf{R}} \frac{1}{m}\sum_{i=1}^m \mathrm{Sign}\left(\langle x_i, u\rangle + \lambda\epsilon_i\right) x_{i,1} - \phi(\lambda)u_1\right]}_{(B)} \\ &+ \underbrace{\mathbb{E}\left[\sup_{\|u-u'\|\leq\epsilon,\lambda\in\mathbf{R}} \frac{1}{m}\sum_{i=1}^m \left(\mathrm{Sign}\left(\langle x_i, u\rangle + \lambda\epsilon_i\right) - \mathrm{Sign}\left(\left\langle x_i, u'\right\rangle + \lambda\epsilon_i\right)\right) x_{i,1} + \phi(\lambda)\left(u_1' - u_1\right)\right]}_{(C)}. \end{aligned} \tag{124}$$

Here $\phi(\lambda) = \sqrt{\frac{2}{\pi}}(1-p) + \sqrt{\frac{2}{\pi}}p\mathbb{E}\left[e^{-\lambda^2\epsilon^2/2}\right]$ is the same as before. We first control (B). To this goal, we show that $\sup_{\lambda\in\mathbf{R}} \frac{1}{m}\sum_{i=1}^m \operatorname{Sign}\left(\langle x_i, u\rangle + \lambda\epsilon_i\right) x_{i,1} - \phi(\lambda)u_1$ is $\mathcal{O}(1/m)$-sub-Gaussian. We prove it via checking the sub-Gaussian norm

$$\left\|\sup_{\lambda\in\mathbf{R}} \operatorname{Sign}\left(\langle x_i, u\rangle + \lambda\epsilon_i\right) x_{i,1} - \phi(\lambda)u_1\right\|_{\psi_2} \le \left\||x_{i,1}|\right\|_{\psi_2} + \sqrt{\frac{2}{\pi}} = \mathcal{O}(1). \tag{125}$$

Hence, via maximum inequality, we have

$$(B) \le \underbrace{\mathbb{E}\left[\sup_{\lambda\in\mathbf{R}} \frac{1}{m}\sum_{i=1}^m \operatorname{Sign}\left(\langle x_i, u\rangle + \lambda\epsilon_i\right) x_{i,1} - \phi(\lambda)u_1\right]}_{(D)} + \mathcal{O}\left(\sqrt{\frac{d\log\left(\frac{1}{\epsilon}\right)}{m}}\right). \tag{126}$$

To control (D), we further decompose it into two parts,

$$\begin{aligned}(D) \le\ & \underbrace{\mathbb{E}\left[\sup_{\nu\in[0,1]} \frac{1}{m}\sum_{i=1}^m \operatorname{Sign}\left(\nu\left\langle x_i, u\right\rangle + \epsilon_i\right) x_{i,1} - \phi\left(\frac{1}{\nu}\right)u_1\right]}_{(D_1)} \\ & + \underbrace{\mathbb{E}\left[\sup_{\lambda\in[0,1]} \frac{1}{m}\sum_{i=1}^m \operatorname{Sign}\left(\langle x_i, u\rangle + \lambda\epsilon_i\right) x_{i,1} - \phi(\lambda)u_1\right]}_{(D_2)}.\end{aligned} \tag{127}$$

To control (D_1) and (D_2) we use arguments based on bracketing maximal inequality. We defer a preliminary discussion on bracketing maximal inequality to Appendix G. We first control (D_1). Let $\mathbb{T}_\xi$ be defined as the ξ-net of the interval $[0,1]$. We show that for any $\nu, \nu' \in [0,1]$ such that $|\nu - \nu'| \le \xi$, we can control $\left\|\left(\operatorname{Sign}(\nu\left\langle x_i, u\right\rangle + \epsilon_i) - \operatorname{Sign}(\nu'\left\langle x_i, u\right\rangle + \epsilon_i)\right) x_{i,1}\right\|_{L_2(\mathbb{P})}$. To this goal, we first have

$$\begin{aligned}&\mathbb{E}\left[\left(\operatorname{Sign}(\nu\left\langle x_i, u\right\rangle + \epsilon_i) - \operatorname{Sign}(\nu'\left\langle x_i, u\right\rangle + \epsilon_i)\right)^2 x_{i,1}^2\right] \\ &\lesssim \mathbb{E}\left[\left|\operatorname{Sign}(\nu\left\langle x_i, u\right\rangle + \epsilon_i) - \operatorname{Sign}(\nu'\left\langle x_i, u\right\rangle + \epsilon_i)\right|\right] \\ &\le \mathbb{E}\left[\mathbb{1}\left(\left|\left(\nu - \nu'\right)\left\langle x_i, u\right\rangle\right| \ge t\right) + \mathbb{1}\left(\left|\nu\left\langle x_i, u\right\rangle + \epsilon_i\right| \le t\right)\right] \\ &\lesssim e^{-C\frac{t^2}{\xi^2}} + t.\end{aligned} \tag{128}$$

Upon picking $t \asymp \xi\log\left(\frac{1}{\xi}\right)$, we have

$$\left\|\left(\operatorname{Sign}(\nu\left\langle x_i, u\right\rangle + \epsilon_i) - \operatorname{Sign}(\nu'\left\langle x_i, u\right\rangle + \epsilon_i)\right) x_{i,1}\right\|_{L_2(\mathbb{P})} \lesssim \sqrt{\xi\log\left(\frac{1}{\xi}\right)}. \tag{129}$$

Therefore, the bracketing number is bounded by $N_{[]}(\varepsilon\|F\|, \mathcal{F}, \|\cdot\|) \lesssim C\frac{1}{\sqrt{\epsilon}}$, which in turn leads to an upper bound on the bracketing entropy $J_{[]}(1, \mathcal{F}, L_2(\mathbb{P})) \lesssim 1$. Applying Theorem 6 leads to

$$(D_1) \lesssim \sqrt{\frac{1}{m}}. \tag{130}$$

Similarly, we can show that $(D_2) \lesssim\sqrt{\frac{1}{m}}$. Therefore, we conclude that

$$(B) \lesssim\sqrt{\frac{d\log\left(\frac{1}{\epsilon}\right)}{m}}. \tag{131}$$

For (C), we can use the similar technique in the overparameterized setting $(m \ll d)$, which leads to

$$(C) \lesssim\sqrt{\log(m)}\epsilon +\sqrt{\frac{d\log(m)}{m}}. \tag{132}$$

Therefore, once we set $\epsilon \asymp\sqrt{\frac{d}{m}}$, we immediately obtain

$$(A) \lesssim\sqrt{\frac{d\log(m)}{m}}. \tag{133}$$

Combining the derived bounds results in

$$\mathbb{P}\left(\sup_{u\in\mathbf{R}^d}\left\|\frac{1}{m}\sum_{i=1}^{m}\text{Sign}\left(\langle x_i,u\rangle+\epsilon_i\right)x_i-\varphi(u)\frac{u}{\|u\|}\right\|_\infty \geq C\sqrt{\frac{d\log(m)}{m}}+\delta\right)\leq e^{c_1\log(d)-c_2m\delta^2}. \tag{134}$$

Assuming $m \gtrsim \frac{d\log(m)}{(1-p)^2}$, the above bound reduces to

$$\mathbb{P}\left(\sup_{u\in\mathbf{R}^d}\left\|\frac{1}{\varphi(u)}\frac{1}{m}\sum_{i=1}^{m}\text{Sign}\left(\langle x_i,u\rangle+\epsilon_i\right)x_i-\frac{u}{\|u\|}\right\|_\infty \geq \delta\right)\leq e^{-cm\delta^2}. \tag{135}$$

E Auxiliary Lemmas

Lemma 6. *Suppose $x_1,\cdots,x_m$ are i.i.d. standard Gaussian vectors with dimension d. Then, for arbitrary $\delta>0$ we have*

$$\mathbb{P}\left(\sup_{\|u\|=1}\left|\frac{1}{m}\sum_{i=1}^{m}|\langle x_i,u\rangle|-\sqrt{\frac{2}{\pi}}\right|\geq C\sqrt{\frac{d}{m}}+\delta\right)\leq e^{-cm\delta^2}. \tag{136}$$

Here C,c are universal constants.

Proof. This lemma directly follows from the standard expectation and high probability bounds for sub-Gaussian process. See e.g., [30, Lemma 4] for a simple proof. □

Lemma 7. *For two arbitrary vectors $a,b\in\mathbb{R}^n$, we have*

$$\|a\odot b\|\leq\|a\|_\infty\|b\|. \tag{137}$$

F Deferred Proofs

F.1 Proof of Lemma 5

Proof. To prove this lemma, it suffices to show that, for any $u, v \in \mathbf{R}^d$, we have

$$\mathbb{E}\left[\operatorname{Sign}\left(\epsilon + \langle x, u\rangle\right)\langle x, v\rangle\right] = \sqrt{\frac{2}{\pi}}\mathbb{E}\left[e^{-\epsilon^2/2\|u\|^2}\right]\left\langle \frac{u}{\|u\|}, v\right\rangle. \tag{138}$$

Without loss of generality, we assume that $\|u\| = \|v\| = 1$. Let us denote $w := \langle x, u\rangle, z := \langle x, v\rangle, \rho := \operatorname{Cov}(w, z) = \langle u, v\rangle$. Then

$$\begin{aligned}
\mathbb{E}\left[\operatorname{Sign}\left(\epsilon + \langle x, u\rangle\right)\langle x, v\rangle\right] &= \mathbb{E}\left[\operatorname{Sign}\left(\epsilon + w\right) z\right] \\
&\overset{(a)}{=} \rho\mathbb{E}\left[\operatorname{Sign}(w + \epsilon) w\right] \\
&= \rho\mathbb{E}_\epsilon\left[\int_{-\epsilon}^{\infty} t\frac{1}{\sqrt{2\pi}}e^{-t^2/2}dt - \int_{-\infty}^{-\epsilon} t\frac{1}{\sqrt{2\pi}}e^{-t^2/2}dt\right] \\
&= \rho\mathbb{E}_\epsilon\left[\int_{-\epsilon}^{\infty} t\frac{1}{\sqrt{2\pi}}e^{-t^2/2}dt + \int_{\epsilon}^{\infty} t\frac{1}{\sqrt{2\pi}}e^{-t^2/2}dt\right] \\
&= 2\rho\mathbb{E}_\epsilon\left[\int_{|\epsilon|}^{\infty} t\frac{1}{\sqrt{2\pi}}e^{-t^2/2}dt\right] \\
&= \sqrt{\frac{2}{\pi}}\langle u, v\rangle\mathbb{E}_\epsilon\left[\int_{|\epsilon|}^{\infty} d\left(-e^{-t^2/2}\right)\right] \\
&= \sqrt{\frac{2}{\pi}}\langle u, v\rangle\mathbb{E}_\epsilon\left[e^{-\epsilon^2/2}\right].
\end{aligned} \tag{139}$$

Here in (a) we use the fact that $z|w, \epsilon \sim \mathcal{N}(\rho w, 1 - \rho^2)$ since ϵ is independent of w, z. Hence, we have

$$\mathbb{E}\left[\operatorname{Sign}\left(\epsilon + \langle x, u\rangle\right)\langle x, v\rangle\right] = \sqrt{\frac{2}{\pi}}\mathbb{E}\left[e^{-\epsilon^2/2\|u\|^2}\right]\left\langle \frac{u}{\|u\|}, v\right\rangle \tag{140}$$

for any $u, v \in \mathbf{R}^d$. On the other hand, it is easy to verify that $\mathbb{E}\left[\operatorname{Sign}\left(\langle x, u\rangle\right)\langle x, v\rangle\right] = \sqrt{\frac{2}{\pi}}\left\langle \frac{u}{\|u\|}, v\right\rangle$. The proof is completed by noting that the corruption probability is p. □

G Preliminaries on the Uniform Concentration Bounds

In this section, we provide the preliminary probability tools for proving Proposition 1.

Definition 2 (Sub-Gaussian random variable)**.** *We say a random variable $X \in \mathbf{R}$ with expectation $\mathbb{E}[X] = \mu$ is σ^2-sub-Gaussian if for all $\lambda \in \mathbf{R}$, we have $\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq e^{\frac{\lambda^2\sigma^2}{2}}$. Moreover, the sub-Gaussian norm of X is defined as $\|X\|_{\psi_2} := \sup_{p\geq 1}\left\{p^{-1/2}(\mathbb{E}[|X|^p])^{1/p}\right\}$.*

According to [41], the following statements are equivalent:

- X is σ^2-sub-Gaussian.

- (Tail bound) For any $t > 0$, we have $\mathbb{P}(|X - \mu| \geq t) \leq 2e^{-\frac{t^2}{2\sigma^2}}$.
- (Moment bound) We have $\|X\|_{\psi_2} \lesssim \sigma$.

Next, we provide the definitions of the sub-Gaussian process, ϵ-net, and covering number.

Definition 3 (Sub-Gaussian process)**.** *A zero mean stochastic process $\{\mathcal{X}_\theta, \theta \in \mathbb{T}\}$ is a σ^2-sub-Gaussian process with respect to a metric d on a set $\mathbb{T}$, if for every $\theta, \theta' \in \mathbb{T}$, the random variable $\mathcal{X}_\theta - \mathcal{X}_{\theta'}$ is $(\sigma d(\theta, \theta'))^2$-sub-Gaussian.*

Definition 4 (ϵ-net and covering number)**.** *A set $\mathcal{N}$ is called an ϵ-net for $(\mathbb{T}, d)$ if for every $t \in \mathbb{T}$, there exists $\pi(t) \in \mathcal{N}$ such that $d(t, \pi(t)) \leq \epsilon$. The covering number $N(\mathbb{T}, d, \epsilon)$ is defined as the smallest cardinality of an ϵ-net for $(\mathbb{T}, d)$:*

$$N(\mathbb{T}, d, \epsilon) := \inf\{|\mathcal{N}| : \mathcal{N} \text{ is an } \epsilon\text{-net for } (\mathbb{T}, d)\}.$$

Definition 5 (Bracketing number, Definition 2.1.6 in [38])**.** *Given two functions l and u, the bracket $[l, u]$ is the set of all functions f with $l \leq f \leq u$. An ε-bracket is a bracket $[l, u]$ with $\|u - l\| < \varepsilon$. The bracketing number $N_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|)$ is the minimum number of ε-brackets needed to cover $\mathcal{F}$. The bracketing entropy is the logarithm of the bracketing number. In the definition of the bracketing number, the upper and lower bounds u and l of the brackets need not belong to $\mathcal{F}$ themselves but are assumed to have finite norms.*

Bracketing number can be regarded as an analog of covering number, describing the geometric complexity of the underlining function space. Although bracketing number of a general function class is difficult to characterize, for some specific function classes, we can easily derive upper bounds for their bracketing number. In particular, we have the following result for Lipschitz functions.

Theorem 5 (Adapted from Theorem 2.7.11 in [38])**.** *Let $\mathcal{F} = \{f_t : t \in T\}$ be a class of functions. Suppose that for arbitrary $s, t \in T$, we have*

$$|f_s(x) - f_t(x)| \leq d(s, t) F(x), \tag{141}$$

for some metric d on the index set, function F on the sample space, and every x. Then, for any norm $\|\cdot\|$,

$$N_{[]}(2\varepsilon\|F\|, \mathcal{F}, \|\cdot\|) \leq N(\varepsilon, T, d). \tag{142}$$

Theorem 6 (Adapted from Theorem 2.14.2 in [38])**.** *For a given norm $\|\cdot\|$, define a bracketing integral of a class of functions $\mathcal{F}$ as*

$$J_{[]}(\delta, \mathcal{F}, \|\cdot\|) = \int_0^\delta \sqrt{1 + \log N_{[]}(\varepsilon\|F\|, \mathcal{F}, \|\cdot\|)} d\varepsilon. \tag{143}$$

Let $\mathcal{F}$ be a class of measurable functions with measurable envelope function F, we have

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}[f(X)]\right] \lesssim J_{[]}(1, \mathcal{F}, L_2(\mathbb{P})) \frac{\|F\|_{L_2(\mathbb{P})}}{\sqrt{n}}, \tag{144}$$

where $\mathbb{P}$ is the distribution of X, and the $L_2(\mathbb{P})$-norm is defined as $\|f\|_{L_2(\mathbb{P})} := \left(\int_\Omega f^2(\omega) d\mathbb{P}(\omega)\right)^{1/2}$.