The Implicit Regularization of Dynamical Stability in Stochastic Gradient Descent

Lei Wu [1 2] **Weijie J. Su** [3]

Abstract

In this paper, we study the implicit regularization of stochastic gradient descent (SGD) through the lens of *dynamical stability* (Wu et al., 2018). We start by revising existing stability analyses of SGD, showing how the Frobenius norm and trace of Hessian relate to different notions of stability. Notably, if a global minimum is linearly stable for SGD, then the trace of Hessian must be less than or equal to $2/\eta$, where η denotes the learning rate. By contrast, for gradient descent (GD), the stability imposes a similar constraint but only on the largest eigenvalue of Hessian. We then turn to analyze the generalization properties of these stable minima, focusing specifically on two-layer ReLU networks and diagonal linear networks. Notably, we establish the *equivalence* between these metrics of sharpness and certain parameter norms for the two models, which allows us to show that the stable minima of SGD provably generalize well. By contrast, the stability-induced regularization of GD is provably too weak to ensure satisfactory generalization. This discrepancy provides an explanation of why SGD often generalizes better than GD. Note that the learning rate (LR) plays a pivotal role in the strength of stability-induced regularization. As the LR increases, the regularization effect becomes more pronounced, elucidating why SGD with a larger LR consistently demonstrates superior generalization capabilities. Additionally, numerical experiments are provided to support our theoretical findings.

[1]School of Mathematical Sciences, Peking University, Beijing, China [2]Center for Machine Learning Research, Peking University, Beijing, China [3]Wharton Statistics and Data Science Department, University of Pennsylvania, Philadelphia, USA. Correspondence to: Lei Wu <leiwu@math.pku.edu.cn>, Weijie J. Su <suw@wharton.upenn.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

1. Introduction

In modern machine learning, models are often over-parameterized in the sense that they can easily interpolate all training data. Therefore, one may be concerned that algorithms may pick up solutions that generalize badly on test data (Wu et al., 2017). Fortunately, it has been found that simple SGD and its variants always converge to solutions that generalize well, even without empolying any explicit regularizations (Zhang et al., 2017). Furthermore, SGD often generalizes better than GD (Keskar et al., 2017). Hence, there must exist certain "implicit regularization" mechanisms at work (Neyshabur et al., 2014). As practitioners increasingly rely on implicit regularization to mitigate overfitting, it becomes imperative to understand the underlying mechanisms.

The most popular explanation is the *flat-minima hypothesis*: SGD tends to select flat minima (Keskar et al., 2017) and flat minima generalize well (Hochreiter & Schmidhuber, 1994; 1997). This hypothesis has been widely adopted in practice to tune the hyperparameters of SGD (Keskar et al., 2017; Jastrzębski et al., 2017; Wu et al., 2020b) and to design new optimizers (Izmailov et al., 2018; Foret et al., 2020; Wu et al., 2020a) for better generalization. Despite its widespread use, the theoretical understanding is still largely lacking: 1) Why does SGD favor flat minima? 2) Why do flat minima generalize?

In this paper, we aim to address these questions by adopting the perspective of dynamical stability (Wu et al., 2018; 2022). For over-parameterized models, all global minima are fixed points of SGD but their stability can be different. Notably, when confronted with a small perturbation, SGD steers away from unstable minima, while stable minima tend to be more resilient, allowing SGD to persist and even reconverge after initial perturbations. This intriguing behavior suggests that SGD exhibits a preference for stable minima. The remaining puzzle lies in understanding the relationship between the stability of a minimum, its sharpness, and its generalization properties.

It is well-known that the stability condition for GD is $\|H(\theta)\|_2 \leqslant 2/\eta$ (Wu et al., 2018), where $H(\cdot)$ denotes the Hessian matrix. This implies that GD tends to select

minima whose sharpness, as measured by the spectral norm of Hessian, is bounded independently of the model size and sample size. Mulayoff et al. (2021); Nacson et al. (2022) showed that for univariate two-layer ReLU networks and diagonal linear networks, this sharpness can control the model capacity under some data assumption. Therefore, the stability ensures that GD selects flat minima that generalize well for these models.

Then a natural question is: *Can we establish a similar understanding of SGD?* Ma & Ying (2021) showed that if a global minimum is *linearly stable* for SGD, then the trace of Hessian $\mathrm{Tr}(H(\theta))$ must be bounded. Meanwhile, it was also proved that for ReLU networks, $\mathrm{Tr}(H(\theta))$ can control the Sobolev seminorm of the functions implemented. These together provide insight into how dynamical stability can act as a form of regularization in SGD. See Wu et al. (2017) for a similar argument. However, it is important to note that this smoothness-based generalization cannot explain the superiority of neural networks in high dimensions (Barron, 1993) as the resulting generalization error bound suffers from the curse of dimensionality. The major reason is that the upper bound of $\mathrm{Tr}(H(\theta))$ obtained in Ma & Ying (2021) grows linearly with the number of parameters. In contrast, by introducing a new notion of stability, Wu et al. (2022) showed that the stability imposes a size-independent control on the Frobenius norm of Hessian: $\|H(\theta)\|_F$ but Wu et al. (2022) did not discuss the corresponding generalization properties. In a word, understanding the stability-induced regularization is still incomplete for SGD and in particular, the following critical questions remain to be answered:

- Can we show that the stable minima of SGD generalize well in high dimensions?
- Can we explain why SGD generalizes better than GD?

Our contributions. We begin by presenting an improved stability analysis for SGD, demonstrating that stability imposes a size-independent control on either the Frobenius norm or the trace of the Hessian matrix, depending on the notion of stability used. Specifically, if a global minimum θ is *linearly stable* for SGD, then $\mathrm{Tr}(H(\theta)) \leqslant 2/\eta$; if it satisfies a loss stability, then $\|H(\theta)\|_F = O(1/\eta)$. In contrast, the stability of GD only controls the largest eigenvalue of Hessian: $\|H(\theta)\|_2 \leqslant 2/\eta$. We then examine the implications of these stability conditions for generalization, and our main findings are summarized as follows.

- We first consider two-layer ReLU networks. It is proved that all three aforementioned measures of sharpness can effectively bound the path norm (Neyshabur et al., 2015; E et al., 2019), thus controlling the generalization gap. As a result, for both SGD and GD, the stable minima are guaranteed to generalize well, which stems from the stability conditions that impose constraints ensuring that the path norms remain bounded by $O(1/\eta)$, irrespective of the model's size. Thus, the size-independent nature of sharpness control strengthens the assurance of favorable generalization properties for the stable minima.
- We next delve into the analysis of diagonal linear networks, which are essentially over-parameterized linear models. We prove that the spectral norm, Frobenius norm, and trace of Hessian are roughly equivalent to the ℓ_∞, ℓ_2 and ℓ_1 norm of the effective coefficients, respectively. The stability of GD only guarantees a size-independent control on the spectral norm of Hessian, thereby the ℓ_∞ norm of effective coefficients. Consequently, stable minima of GD may not generalize well since the ℓ_∞ norm cannot yield an effective capacity control for linear models. In stark contrast, the stability of SGD imposes size-independent controls on the trace or Frobenius norm of Hessian, thereby the ℓ_1 or ℓ_2 norm of effective coefficients. As a result, the stable minima of SGD must generalize well.

  This comparison between SGD and GD effectively demonstrates that in the case of diagonal linear networks, the stability of SGD imparts a substantially stronger regularization effect than that of GD. This provides an explanation for the superior generalization performance consistently observed in SGD over GD.

It is important to note that the strength of stability-induced regularization crucially depends on the size of LR. A larger LR imposes a stricter constraint on the sharpness of stable minima, thereby enforcing SGD/GD to select flatter minima. This explains why SGD with a large LR often generalizes better. To support our theoretical findings, systematic numerical experiments are provided and in particular, we examine in detail the impact of varying LRs.

1.1. Related works

Implicit regularization of SGD. In SGD, there exist multiple mechanisms that contribute to the implicit regularization (Su, 2021; He & Su, 2020; Vardi, 2022). One is the specific dynamical process, along with small initialization, aiding SGD in finding solutions that generalize well (Zhang et al., 2017; Woodworth et al., 2020; Blanc et al., 2020; Chizat & Bach, 2020; Pesme et al., 2021; Ma et al., 2020; Xu et al., 2021). This type of implicit regularization heavily relies on the initialization size. In contrast, the stability-induced regularization (Wu et al., 2018) is independent of the initialization and can explain why using a large LR and small batch size is more favorable (Wu et al., 2022; Ma & Ying, 2021). In the experiments of the current work, we intentionally exclude the small initialization-induced regularization by using a large initialization, as our focus is

understanding the stability-induced regularization.

Barrett & Dherin (2020); Smith et al. (2020) explained the benefit of using a large LR through a modified equation analysis. However, the analysis is only validated for a finite time and hence, cannot explain why SGD favors certain minima, as the latter is a long-time property. In contrast, our stability analysis does not have this limitation.

Generalization of flat minima. To explain why flat minima generalize well, many works rely on the PAC-Bayesian argument (McAllester, 1999). This argument established the connection between certain sharpness and the average generalization error of perturbed solutions, which, however, is not for the original one (Neyshabur et al., 2017; Tsuzuku et al., 2020). Furthermore, Bayesian arguments tend to ignore the specific parametrization of neural networks. (Mulayoff et al., 2021; Ma & Ying, 2021; Nacson et al., 2022) established sharpness-based generalization bounds for some neural networks but they are limited to either linear or low-dimensional cases. In contrast, our sharpness-based generalization bound of two-layer ReLU networks is effective in high dimensions.

In addition, some works argue that sharpness itself can not effectively control model capacity since ReLU neural nets are invariant to node-wise rescaling, whereas the sharpness is not. Consequently, sharp minima can generalize well (Dinh et al., 2017). To overcome this issue, various rescaling-invariant sharpness has been proposed, e.g., the Fisher-Rao metric (Liang et al., 2019), normalized flatness (Tsuzuku et al., 2020), relative flatness (Petzka et al., 2021). However, our analysis suggests that flatness can be *sufficient* for good generalization.

Notation. For an integer k, let $[k] = \{1, 2, \dots, k\}$. For a vector v, let $\|v\|_p = (\sum_i v_i^p)^{1/p}$, $\|\cdot\| = \|\cdot\|_2$, and $\hat{v} = v/\|v\|_2$. For a matrix A, denote by $\|A\|_2$ and $\|A\|_F$ the spectral norm and the Frobenius norm, respectively and let $\{\lambda_i(A)\}_{i\geqslant 1}$ be the eigenvalues of A in a decreasing order. Let $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d \,|\, \|x\| = 1\}$ and $r\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d \,|\, \|x\| = r\}$. For a distribution μ, let $\|f\|^2_{L_2(\mu)} = \mathbb{E}_{x\sim\mu}[f^2(x)]$. We will use C to denote an absolute constant, whose value may change from line to line. For notation simplicity, we write $X \lesssim Y$ if $X \leqslant CY$ and $X \gtrsim Y$ if $X \geqslant CY$. Analogously, we write $X \sim Y$ if $X \lesssim Y$ and $X \gtrsim Y$ hold simultaneously.

2. Preliminaries

Let $S = \{(x_i, y_i = f^*(x_i))\}_{i=1}^n$ be the training set, where $x_1, \dots, x_n$ are i.i.d. samples drawn from the input distribution ρ and $f^* : \mathbb{R}^d \mapsto \mathbb{R}$ be the target function. Our task is to recover f^* from S. Let $f(\cdot;\theta) : \mathbb{R}^d \mapsto \mathbb{R}$ be our model parameterized by $\theta \in \mathbb{R}^p$, where d and p denote the input dimension and the model size (i.e., the number of parameters), respectively. The empirical and population risk are given by

$$\begin{aligned} \hat{\mathcal{R}}(\theta) &= \frac{1}{2n}\sum_{i=1}^n (f(x_i;\theta) - y_i)^2 \\ \mathcal{R}(\theta) &= \frac{1}{2}\,\mathbb{E}_{x,y}[(f(x;\theta) - y)^2], \end{aligned} \tag{1}$$

where the square loss is used. Throughout this paper, we make the following over-parameterization assumption.

Assumption 2.1 (Over-parameterization)**.** $\min_\theta \hat{\mathcal{R}}(\theta) = 0$

Let $g_i(\theta) = \nabla f(x_i;\theta)$ and $e_i(\theta) = f(x_i;\theta) - y_i$. Then the Hessian matrix is given by

$$H(\theta) = \frac{1}{n}\sum_{i=1}^n g_i(\theta) g_i(\theta)^T + \frac{1}{n}\sum_{i=1}^n e_i(\theta)\nabla^2 f(x_i;\theta). \tag{2}$$

Let $G(\theta) = \frac{1}{n}\sum_{i=1}^n g_i(\theta) g_i(\theta)^T$ be the associate empirical Fisher matrix. Then, (2) implies that when fitting errors are small, we have $H(\theta) \approx G(\theta)$ and in particular, $H(\theta) = G(\theta)$ if θ is a global minimum. Note that $G(\theta)$ is always positive semi-definite but $H(\theta)$ is not. In our analysis of the dynamical stability, we shall focus on the region with small empirical risk and hence, we do not distinguish the Fisher matrix and Hessian matrix too much since they are close to each other.

Gradient clipping. In our experiments, we will use large initialization to exclude the implicit regularization induced by small initialization. This choice will make it very often that SGD and GD with a large LR diverge initially although there exist stable minima on landscape. To resolve this issue, we shall apply gradient clipping (Pascanu et al., 2013; Mikolov et al., 2012) to stabilize the training. Specifically, we use the following clipped (stochastic) gradient for SGD/GD update:

$$\nabla\hat{\mathcal{R}}_{clip}(\theta) = \min\{\|\nabla\hat{\mathcal{R}}(\theta)\|, \delta\}\frac{\nabla\hat{\mathcal{R}}(\theta)}{\|\nabla\hat{\mathcal{R}}(\theta)\|},$$

where δ denotes the clipping threshold. In all our experiments, we find that gradient clipping is activated only during the early and intermediate training stages, and will be automatically switched off when SGD/GD nearly converges since the gradient norm there is lower than the clipping threshold. Therefore, gradient clipping does not change the dynamical stability of SGD/GD at global minima.

3. The dynamical stability of SGD

In this section, we consider three measures of sharpness: $\|G(\theta)\|_2$, $\|G(\theta)\|_F$, and $\mathrm{Tr}(G(\theta))$ and study how they are related to the stability of SGD and GD

It is well-known that if θ is a linearly stable for GD, then $\|H(\theta)\|_2 \leqslant 2/\eta$ (Wu et al., 2018; Mulayoff et al., 2021),

which implies $\|G(\theta)\|_2 \leqslant 2/\eta$ if θ is a global minima. Next, we will show that similar size-independent controls hold for SGD but on different norms of Fisher matrix.

3.1. Linear stability

Consider the mini-batch SGD:

$$\theta_{t+1} = \theta_t - \eta(f(x_{i_t};\theta_t) - y_{i_t})\nabla f(x_{i_t};\theta_t), \quad (3)$$

where $i_t \overset{iid}{\sim} \mathrm{Unif}([n])$. Throughout this paper, we assume the batch size to be 1 for simplicity.

Suppose that θ_t converges to a global minimum θ^*. Let $\delta_t = \theta_t - \theta^*$ be the deviation. When $\|\delta_t\|$ is small, $f(x;\theta_t) = f(x;\theta^*) + \nabla f(x;\theta^*)^T\delta_t + o(\|\delta_t\|)$. Substituting it into (3) and noticing $y_i = f(x_i;\theta^*)$, we obtain the linearized SGD:

$$\delta_{t+1} = \delta_t - \eta\nabla f(x_i;\theta^*)\nabla f(x_i;\theta^*)^T\delta_t, \quad (4)$$

where the high-order term is neglected. This linearized SGD characterizes how δ_t evolves when θ_t is close to θ^*.

Definition 3.1 (Linear stability)**.** Let $(\delta_t)_{t\in\mathbb{N}}$ be the solution of the linearized SGD (4). A global minimum θ^* is said to be linearly stable if $\|\mathbb{E}[\delta_t\delta_t^T]\|_F \leqslant \|\mathbb{E}[\delta_0\delta_0^T]\|_F$ for any $t \in \mathbb{N}$ and initial distribution over δ_0.

The linear stability defined above measures the instability by using the second-order moment of deviations. If θ^* is not linearly stable, it is unlikely that θ_t converges to θ^*. The following provides a necessary condition of linear stability, whose proof can be found in Appendix B.1.

Proposition 3.2. *If a global minimum θ^* is linearly stable, then* $\mathrm{Tr}(G(\theta^*)) \leqslant 2/\eta$.

This proposition implies that SGD tends to select minima, where the sharpness–as measured by the trace of Hessian– is bounded by $2/\eta$, independently of the model size and sample size. This size-independence means that stability imposes an effective sharpness control no matter how over-paramterized the model is and this in turn will yield an effective control on the model capacity as demonstrated in our subsequent generalization analysis. In contrast, for GD, the linear stability imposes a much weaker control: Only the largest eigenvalue of Hessian is bounded by $2/\eta$.

Comparison with existing works. Défossez & Bach (2015) derived the same upper bound of learning rate for least square problems and studied its impact on the convergence of SGD. In contrast, our focus is understanding the implication for regularization. Moreover, it should be stressed that our stability condition is derived by examing the linearized SGD but relevant for nonlinear SGD. Ma & Ying (2021) also derived an upper bound of $\mathrm{Tr}(G(\theta))$ by examing the linear stability but their bound grows explicitly with the model size. Specifically, Ma & Ying (2021, Theorem 2) gives the bound $\mathrm{Tr}(G) \leqslant 2p/\eta$, where p is the number of parameters.

3.2. Loss stability

In this section, we revise the *loss stability* defined in Wu et al. (2022), which is applicable to a general SGD:

$$\theta_{t+1} = \theta_t - \eta(\nabla\hat{\mathcal{R}}(\theta_t) + \xi_t), \quad (5)$$

where ξ_t denotes a general gradient noise that satisfies

$$\mathbb{E}[\xi_t] = 0, \quad \Sigma(\theta_t) := \mathbb{E}[\xi_t\xi_t^T] = 2\hat{\mathcal{R}}(\theta_t)S(\theta_t). \quad (6)$$

Here $S(\theta)$ represents the loss-scaled noise covariance matrix. This assumption of gradient noise implies that the noise magnitude is proportional to the loss value, which is naturally satisfied by the mini-batch SGD (3) as pointed out in Mori et al. (2022); Wu et al. (2022); Wojtowytsch (2021); Feng & Tu (2021); Liu et al. (2021).

Lemma 3.3 (One-step update)**.** *Suppose $\hat{\mathcal{R}} \in C^3(\mathbb{R}^p)$. We have* $\mathbb{E}[\hat{\mathcal{R}}(\theta_{t+1})] \geqslant \eta^2\mathrm{Tr}[H(\theta_t)S(\theta_t)]\hat{\mathcal{R}}(\theta_t) + O(\eta^3)$

Proof. By definition, we have

$$\begin{aligned}\hat{\mathcal{R}}(\theta_{t+1}) &= \hat{\mathcal{R}}(\theta_t - \eta\nabla\hat{\mathcal{R}}(\theta_t) - \eta\xi_t)\\ &= \hat{\mathcal{R}}(\theta_t - \eta\nabla\hat{\mathcal{R}}(\theta_t)) + \left\langle \nabla\hat{\mathcal{R}}(\theta_t - \eta\nabla\hat{\mathcal{R}}(\theta_t)), -\eta\xi_t \right\rangle\\ &\quad + \frac{\eta^2}{2}\xi_t^T H(\theta_t - \eta\nabla\hat{\mathcal{R}}(\theta_t))\xi_t + O(\eta^3).\end{aligned}$$

Taking expectation *w.r.t.* ξ_t and using $H(\theta_t - \eta\nabla\hat{\mathcal{R}}(\theta_t)) = H(\theta_t) + O(\eta)$ and $\hat{\mathcal{R}}(\theta_t - \eta\nabla\hat{\mathcal{R}}(\theta_t)) \geqslant 0$ gives

$$\begin{aligned}\mathbb{E}[\hat{\mathcal{R}}(\theta_{t+1})] &\geqslant \frac{\eta^2}{2}\mathrm{Tr}[H(\theta_t)\Sigma(\theta_t)] + O(\eta^3)\\ &= \eta^2\hat{\mathcal{R}}(\theta_t)\mathrm{Tr}[H(\theta_t)S(\theta_t)] + O(\eta^3), \quad (7)\end{aligned}$$

where the last step follows from (6). □

This lemma implies that $\mathrm{Tr}[H(\theta_t)S(\theta_t)]$ determines the local stability if ignoring the higher-order term. Specifically, if the loss $\hat{\mathcal{R}}(\theta_t)$ is sufficiently small such that $H(\theta_t) \approx G(\theta_t)$, then for $\mathbb{E}[\hat{\mathcal{R}}(\theta_{t+1})] \leqslant \hat{\mathcal{R}}(\theta_t)$ to hold, a necessary condition is $\mathrm{Tr}[G(\theta_t)S(\theta_t)] \leqslant 1/\eta^2$. This condition can be converted to a sharpness control by assuming

$$\mu(\theta) := \frac{\mathrm{Tr}(G(\theta)S(\theta))}{\|G(\theta)\|_F^2} \geqslant \mu_0. \quad (8)$$

By treating $G(\theta) \approx H(\theta)$, $\mu(\theta)$ can be interpreted as a factor that quantifies the (loss-scaled) strength of alignment between the noise covariance and local Hessian. For mini-batch SGD (3), Wu et al. (2022) has shown that there exist a size-independent constant $\mu_0 > 0$ such that $\mu(\theta) \geqslant \mu_0$ for neural networks. We refer to Wu et al. (2022) for more discussions on this alignment factor.

Proposition 3.4. *Assume $\hat{\mathcal{R}} \in C^3(\mathbb{R}^p), \mu(\theta) \geqslant \mu_0$. Let $\mathcal{Q}_{\varepsilon,\eta} = \{\theta : \hat{\mathcal{R}}(\theta) \leqslant \varepsilon, \|G(\theta)\|_F > \sqrt{1/\mu_0}/\eta\}$. If $\{\theta_\tau\}_{\tau=0}^t \in \mathcal{Q}_{\varepsilon,\eta}$, then $\mathbb{E}[\hat{\mathcal{R}}(\theta_{t+1})] \geqslant \gamma^t\hat{\mathcal{R}}(\theta_0) + \frac{\gamma^t-1}{\gamma-1}O(\eta^3 + \eta^2\varepsilon^{3/2})$ with $\gamma > 1$.*

The proof can be found in Appendix B.2. This proposition shows that SGD will escape from a low-loss region *exponentially fast* (measured by the loss value) if the the landscape there is too sharp in terms of the Frobenius norm of Hessian. Specifically, SGD can only stay/travel in the region where $\|G(\theta)\|_F \leqslant \sqrt{1/\mu_0}/\eta$.

The above analysis extends Wu et al. (2022) in two aspects. First, our analysis does not need θ_t to be close to a global minimum θ^*, implying that the loss stability is relevant even if SGD has not converge. This is consistent with the numerical experiments in Wu et al. (2022), which shows that the upper bound of Frobenius norm of Hessian matrix holds for the entire training process. Second, our analysis is applicable to general SGD where the gradient noise does not necessarily come from the mini-match sampling. For instance, one can consider the Langevin dynamics with $S(\theta) = G(\theta)$, which has been tested in Zhu et al. (2019) to have similar generalization properties as mini-batch SGD. In contrast, Wu et al. (2022) only considered the mini-batch SGD and required θ^* is close to θ_t in the sense that $\|\theta_t - \theta^*\| = o(1)$.

Remark 3.5. Note that different from the linear stability (Definition 3.1), the loss stability measures the stability by using the changes of loss. In Wu et al. (2022), this stability is referred to as "linear stability". However, based on our preceding explanations, we propose the term "loss stability" as a more fitting term. Furthermore, we will specifically refer to Definition (3.1) as "linear stability", acknowledging that it exclusively holds for the linearized SGD.

3.3. The comparison between two types of stability

The linear stability is defined by examining the linearized SGD (4), which is validated only if θ_t is sufficiently close to a global minimum θ^*. In contrast, the loss stability measures the stability by inspecting if the loss grows exponentially, which is applicable even if θ_t does not converge. In terms of sharpness control, linear stability and loss stability impose size-independent controls on $\mathrm{Tr}(G(\theta))$ and $\|G(\theta)\|_F$, respectively. The former is stronger since $\|G(\theta)\|_F \leqslant \mathrm{Tr}(G(\theta))$. To summarize, linear stability yields a stronger sharpness control but requires the dynamics to be sufficiently close to a global minimum; loss stability is generally relevant but imposes a weaker sharpness control. When the loss-stability is satisfied, SGD may travel in a low-loss region without convergence and the convergence to a global minimum requires the stronger condition of linear stability to be satisfied.

Then a natural question is: Which type of stability characterizes the actual dynamical behavior of SGD better? The answer will depends on the problem and training stages.

- In training practical models, large-LR SGD often takes many iterations to stay in a low-loss region without reaching a global minimum. During these stages, the loss keeps nearly unchanged; and thus, the condition of loss stability must be met but the condition of linear stability is not necessarily to be satisfied. Indeed, the empirical studies by Wu et al. (2022) has demonstrated the relevance of loss stability in this situation.
- In this paper, we focus on detailed analysis of simple models: two-layer ReLU networks and diagonal linear networks, for which we empirically find that linear stability is more relevant. Specifically, the upper bound $2/\eta$ is close to the actual trace of Hessian. While the condition of loss stability is also satisfied, the resulting bound is much looser. These observations are not unexpected as for these simple models, large-LR SGD always converges to zero loss stably, implying the condition of linear stability must be satisfied.

4. Two-layer ReLU networks

We first consider the two-layer ReLU network: $f(x;\theta) = \sum_{j=1}^m a_j \sigma(w_j^T x)$, where $a_j \in \mathbb{R}, w_j \in \mathbb{R}^d$, m denotes the network width, and $\sigma(t) = \max(t, 0)$. In this section, we assume the input distribution to be $\rho = \mathrm{Unif}(\sqrt{d}\mathbb{S}^{d-1})$.

Define the weighted ℓ_2 norm $\|\theta\|_{2,q} := \sum_j (\|w_j\|^2 + q a_j^2)$, where $q > 0$ is the weight factor. The following theorem shows that all three sharpness are equivalent to $\ell_{2,q}$ norms of parameters and the only difference is the weight factor. The proof can be found in Appendix C.2

Theorem 4.1. *For any $\delta \in (0,1)$, let $N(d,\delta) = \inf\{n \in \mathbb{N} : d\log(n/\delta)/n \leqslant 1\}$.*

- *If $n \gtrsim N(d,\delta)$, then w.p. $1-\delta$ we have*
$$\|G(\theta)\|_F \sim \|\theta\|_{2,\sqrt{d}},\ \mathrm{Tr}(G(\theta)) \sim \|\theta\|_{2,d}.$$
- *If $n \gtrsim dN(d,\delta)$, then w.p. $1-\delta$, $\|G(\theta)\|_2 \sim \|\theta\|_{2,1}$.*

Remark 4.2. Sharpness is a data-dependent quantity since it measures the local curvature of *empirical landscape*. In contrast, a weighted ℓ_2 norm of parameters is data independent. The equivalence shown in Theorem 4.1 is possible because we assume ρ to be isotropic. A question of more interest would be to exploit the effect of data dependence by making an anisotropic assumption on ρ, which we leave to the future work.

For ReLU networks, it is well-known that the generalization gap can be controlled by the path norm (Neyshabur et al., 2015; E et al., 2021) $\|\theta\|_{\mathcal{P}} := \sum_j |a_j| \|w_j\|$. By the AM-GM inequality, we have

$$\|\theta\|_{2,q} = \sum_j (\|w_j\|^2 + q a_j^2)$$

$$\geqslant 2\sqrt{q}\sum_j |a_j|\|w_j\| = 2\sqrt{q}\|\theta\|_{\mathcal{P}}. \qquad (9)$$

This implies that weight ℓ_2 norms can bound the generalization gap although it is not rescaling invariant.

Theorem 4.3. *For SGD and GD with the same LR η, denote by $\hat{\theta}_{sgd}$ and $\hat{\theta}_{gd}$ the linearly stable minimum of SGD and GD, respectively. Suppose* $\sup_{x\in\mathcal{X}}|f^*(x)| \leqslant 1$. *For any $\delta \in (0,1)$, if $n \gtrsim dN(d,\delta)$, then the following holds w.p. at least $1-\delta$*

$$\mathcal{R}(\hat{\theta}_{\rm sgd}) \lesssim \frac{B}{\eta^2 n}, \qquad \mathcal{R}(\hat{\theta}_{\rm gd}) \lesssim \frac{Bd}{\eta^2 n},$$

where $B = \log^3 n + \log(1/\delta)$.

Proof idea. The complete proof can be found in Appendix C.3. Here we provide a sketch of proof idea. For $\hat{\theta}_{\rm sgd}$, the generalization gap can be informally bounded as follows

$$\text{gen-gap}(\hat{\theta}_{\rm sgd}) \overset{(a)}{\lesssim} \frac{d\|\hat{\theta}_{\rm sgd}\|_{\mathcal{P}}^2}{n} \overset{(b)}{\lesssim} \frac{\|\hat{\theta}_{\rm sgd}\|_{2,d}^2}{n} \overset{(c)}{\lesssim} \frac{\mathrm{Tr}^2(G(\hat{\theta}_{\rm sgd}))}{n} \overset{(d)}{\leqslant} \frac{1/\eta^2}{n}, \qquad (10)$$

where (a) follows from the path norm-based generalization bound (Proposition C.16); (b) follows from (9); (c) follows from Theorem 4.1; (d) is due to the stability condition (Proposition 3.2).

This theorem shows that stable minima provably generalize, no matter how over-parameterized the model is. This suggests that the stability-induced regularization is strong enough to eliminate the potential overfitting caused by over-parameterization. In addition, with the same LR, stable minima of SGD generalize better than that of GD. However, it is more fair to compare SGD with LR η and GD with LR $\sqrt{d}\eta$. Thus, we will use this LR choice in our experimental analysis for a fair comparison between SGD and GD.

Comparison with existing works. Mulayoff et al. (2021) conducted a similar analysis for two-layer ReLU networks, which, however, is limited to GD and the univariate case. Another closely related work is Ma & Ying (2021), which established a generalization bound of linearly stable minima of SGD for ReLU networks but the bound suffers from the curse of dimensionality. One of the reasons is that the upper bound of the trace of Hessian derived in Ma & Ying (2021) depends on the model size explicitly. In contrast, our generalization bounds are effective in high dimensions and hold for both SGD and GD.

4.1. Numerical validations

Consider $f^*(x) = \sum_{i=1}^k \sigma(v_i^T x)$ with $v_i \overset{iid}{\sim} \mathrm{Unif}(\mathbb{S}^{d-1})$ and $f(x;\theta) = \sum_{j=1}^m a_j\sigma(w_j^T x)$. We set $k = 10, d = 100, m = 100$, and the sample size $n = 300$. With this choice, the total number of parameters is $p = (d+1)m = 10100$ and thus, we are examining a highly over-parameterized case where $p \gg n$. We consider a large initialization: $a_j \sim \mathcal{N}(0,1)$ and $w_j \sim \mathcal{N}(0, I_d/\sqrt{d})$, with which the path norm at initialization: $\|\theta\|_{\mathcal{P}} \sim m$, growing linearly with the network width. This large initialization excludes the small initialization effect and one must rely on the stability-induced regularization to select minima with small path norms. In addition, gradient clipping will be applied to stabilize the training if SGD/GD blows up initially.

The effect of gradient clipping. Figure 1 shows the dynamical process of SGD with gradient clipping, where $\eta = 1/\sqrt{d}$ and the clipping threshold $\delta = 1$. One can see that the gradient clipping is automatically switched off since around 4000 iterations. After that, SGD can stably converge to a global minimum without clipping operations. This implies that around the convergent minimum, linear stability should be satisfied and consequently, it is not surprising to observe that $\mathrm{Tr}(G(\theta_t)) \leqslant 2/\eta$ when θ_t nearly converge. Another interesting observation is that during the whole training process, $\mathrm{Tr}(G(\theta_t))$ keeps decreasing, which in turn causes the continued decreasing of path norm. This phenomenon cannot be explained by the stability condition and one should delve into the dynamical process of SGD. We refer to Blanc et al. (2020) for a potential explanation.

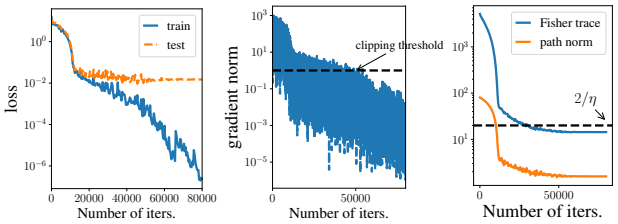

Figure 1. The training process of SGD with gradient clipping. The gradient clipping is automatically switched off in the late phase of training. The trace of Fisher matrix keeps decreasing until it becomes lower than $2/\eta$ and meanwhile, the path norm also keeps decreasing, which is consistent with Theorem 4.1.

The sharpness. Figure 2(a) shows how the sharpness and path norm of minima selected by SGD and GD changes with the LR. For SGD, $\mathrm{Tr}(G(\theta))$ keeps decreasing but close to the upper bound $2/\eta$. Consequently, the path norm also keeps decreasing. This is consistent with the predictions of linear stability analysis and Theorem 4.1. For GD, $\|G(\theta)\|_2$ keeps close to the upper bound $2/\eta$ when the LR is sufficiently large. When the LR is small, the actual sharpness is away from the upper bound. These observations suggest that the impact of stability-induced regularization is particularly significant in the large LR regime.

The test performance. Figure 2(b) shows that the test performance is continually improved for both SGD and GD as increasing the LR, which again confirms the prediction

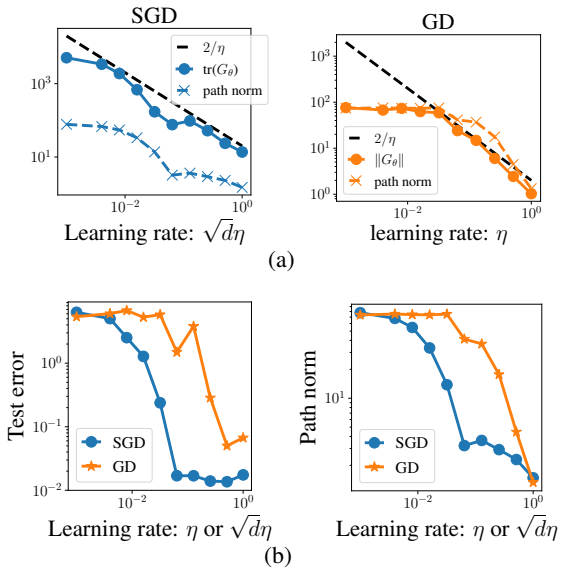

Figure 2. (a) The sharpness and path norm vs. LR. (b) The test performance vs. LR. For a fair comparison, we compare SGD with LR $\eta/\sqrt{d}$ and GD with LR η.

of Theorem 4.3. One observation of more interest is that Figure 2(b) shows that with the fair choice of LR, SGD still generalizes better than GD. This beyond what Theorem 4.3 can explain since the generalization bounds are the same for SGD and GD in such a case. In addition, when the LR is overly large, SGD still generalizes better although its path norm becomes larger than that of GD. A potential explanation is that the noise drives SGD towards better minima with certain mechanism beyond dynamical stability. We leave this to future work.

5. Diagonal linear networks

Consider the two-layer diagonal linear network

$$f(x;\theta) = \langle a \odot b, x \rangle, \tag{11}$$

where $a, b \in \mathbb{R}^d$, $\theta = (a, b) \in \mathbb{R}^{2d}$, and $\odot$ denotes the element-wise multiplication. Despite its simplicity, this model has been widely used in theoretical analysis to demonstrate particular properties of SGD in training neural networks (Woodworth et al., 2020; Gissin et al., 2019; Pesme et al., 2021; Nacson et al., 2022). Note that this model can only represent linear predictors and we will use $\beta = a \odot b$ to denote the effective coefficients. In this section, we make the following assumption on ρ.

Assumption 5.1. Let $X \sim \rho$. Assume $\mathbb{E}[XX^T] = I_d$ and X is sub-Gaussian, i.e., $\|u^T X\|_{\psi_2} \lesssim 1$ for any $u \in \mathbb{S}^{d-1}$.

Here $\|\cdot\|_{\psi_2}$ denotes the sub-Gaussian norm (we refer to Appendix A.3 for details) and one typical example that satisfies the above assumption is $\mathcal{N}(0, I_d)$ and $\mathrm{Unif}([-1,1]^d)$.

Theorem 5.2. *Suppose Assumption 5.1 holds. Let $\alpha = a \odot a + b \odot b$. Let $\delta \in (0,1)$ be the failure probability.*

- *If $r_n = \sqrt{(d + \log(1/\delta))/n} \leqslant 1$, then w.p. $1-\delta$ that*

$$(1 - r_n)\|\alpha\|_\infty \leqslant \|G(\theta)\|_2 \leqslant (1 + r_n)\|\alpha\|_\infty.$$

- *If $\varepsilon_n = \sqrt{\log(d/\delta)/n} \leqslant 1$. Then, w.p. $1-\delta$ that*

$$(1-\varepsilon_n)\|\alpha\|_2 \leqslant \|G(\theta)\|_F \leqslant \varepsilon_n\|\alpha\|_1 + (1+2\varepsilon_n)\|\alpha\|_2$$
$$(1-\varepsilon_n)\|\alpha\|_1 \leqslant \mathrm{Tr}(G(\theta)) \leqslant (1+\varepsilon_n)\|\alpha\|_1.$$

This theorem establishes the equivalence between the sharpness and parameter norms, whose proof is deferred to Appendix D.1. It is worth noting that the cases of Frobenius norm and trace hold in the highly over-parameterized regime: $n \sim \log(d/\delta)$.

Theorem 5.2 shows that with a high probability, $\|G(\theta)\|_2$ is equivalent to $\max_j(a_j^2 + b_j^2)$, which unfortunately cannot provide an effective capacity control for the linear predictor: $(a \odot b)^T x$. Furthermore, the stability of GD only imposes a size-independent control on $\|G(\theta)\|_2$. Thus, we can conclude that the stability-induced regularization of GD is not strong enough to help find generalizable minima. In contrast, the stable minima of SGD provably generalize well, which is explained as follows. Noting

$$\begin{aligned} \|\alpha\|_1 &= \sum_j (a_j^2 + b_j^2) \geqslant 2\sum_j |a_j b_j| = 2\|\beta\|_1 \\ \|\alpha\|_2^2 &= \sum_j (a_j^2 + b_j^2)^2 \geqslant 4\sum_j (a_j b_j)^2 = 4\|\beta\|_2^2 \end{aligned} \tag{12}$$

and applying Theorem 5.2, we can conclude that the linear stability and loss stability can control the ℓ_1 and ℓ_2 norm of effective coefficients, respectively, which yield effective capacity controls for the linear predictor. Specifically, the following theorem formalizes this observation for the case of linear stability and the proof is deferred to Appendix D.2.

Theorem 5.3. *Suppose $\rho = \mathrm{Unif}([-1,1]^d)$ and $f^*(x) = \beta_*^T x$. Let $\hat{\theta} = (\hat{a}, \hat{b})$ be a global minimum that is linearly stable for SGD (3) with LR η. Then, for any $\delta \in (0,1)$, if $n \gtrsim \log(d/\delta)$, then w.p. $1-\delta$ we have $\|\hat{a} \odot \hat{b}\|_1 \lesssim 1/\eta$ and*

$$\mathcal{R}(\hat{\theta}) \lesssim \frac{(1/\eta)^2 \log^2(n)\log(d)}{n} + \frac{(\|\beta_*\|_1 + \frac{1}{\eta})^2 \log(1/\delta)}{n}.$$

This theorem shows that SGD selects minima with the ℓ_1 norm bounded by $1/\eta$. As long as the LR η is sufficiently large and $\|\beta_*\|_1 = O(1)$, the minima found by SGD generalizes well. Woodworth et al. (2020) showed that for this model, gradient flow converges to the minimum ℓ_1 norm solutions when a (near-)zero initialization is used. Nevertheless, we reveal that the linear stability of large-LR SGD has a similar effect, which is independent of the initialization scale.

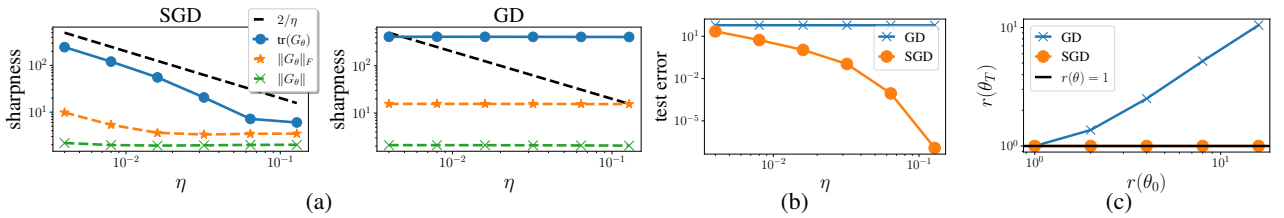

Figure 3. (a) How the sharpness of minima found by SGD and GD changes with the learning rate. We see that for SGD, the upper bound $2/\eta$ provides a quite sharp estimate of the actual trace of Fisher matrix up to a multiplicative constant. In contrast, for GD, the sharpness barely changes as increasing the learning rate. (b) The comparison of test performance between SGD and GD for varying learning rates. (c) Demonstrate the balancing effect of SGD, where the unbalancedness is measured by $r(\theta) = 0.5\|\alpha\|_2/\|\beta\|_1$. The horizontal and vertical axises correspond to the unbalance at initialization and convergence, respectively.

Comparison with Nacson et al. (2022). Nacson et al. (2022) obtained a similar result for GD but it crucially relies on the non-centered data assumption: All coordinates of $\mathbb{E}[X]$ are nonzero. In contrast, our analysis does not need this assumption and moreover, can explain why SGD generalizes better than GD. We also point out that under the non-centered data assumption, the stability-induced regularization might not be able to distinguish SGD and GD as both control the ℓ_1 norm of effective coefficients.

The balancing effect. Another interesting consequence is that SGD tends to select balanced solutions where $a_j^2 \approx b_j^2$ for any $j \in [n]$. This is because minimizing the trace and Frobenius norm of Fisher matrix naturally leads to this balance according to Theorem 5.2 and (12). In contrast, the stability of GD only controls $\max_{j\in[d]}(a_j^2 + b_j^2)$, which does not have the balancing effect except for the coordinate: $k \in \mathrm{argmax}_j(a_j^2 + b_j^2)$.

Deep diagonal linear networks. Similar to Nacson et al. (2022), we can analyze the interaction between depth and stability by examining the deep model $f(x;\theta) = \langle a^D \odot b^D, x\rangle$, where $a^D = (a_1^D, \dots, a_d^D)$ and b^D is defined similarly. Analogous to Theorem 5.2 and (12), one can show that $\mathrm{Tr}(G(\theta)) \gtrsim \|a^D \odot b^D\|_p^p$ with $p = \frac{2(D-1)}{D}$. Thus, the stability-induced regularization changes from the ℓ_1 norm for $D = 2$ to the ℓ_2 norm for $D \to \infty$. Here we do not discuss this in detail since it does not reveal any new insights beyond Nacson et al. (2022, Section 6).

5.1. Numerical validations

Consider $f^*(x) = \beta_*^T x$ with $\beta_* = (1, 1, 1, 0, \cdots, 0)$, for which $\|\beta_*\|_1 = 3$. We set $d = 1000, n = 300$ and initialize the model by $a_j, b_j \overset{iid}{\sim} \mathcal{N}(0, 1)$ for $j = 1, \dots, d$. This large initialization is adopted to eliminate the implicit regularization of small initialization. The model is trained by SGD and GD with varying LRs. Gradient clipping is applied to stabilize the training for the case of large LR. The results are reported in Figure 3.

The sharpness. The left panel of Figure 3(a) shows how the actual sharpness of minima selected by SGD changes as increasing the LR. One can see that the $\mathrm{Tr}(G(\theta))$ keeps close to $2/\eta$–the upper bound ensured by the linear stability; $\|G(\theta)\|_F$ also decreases with LR though the decreasing is not significant. These are in contrast to $\|G(\theta)\|_2$, which keeps almost unchanged. These observations suggest that for diagonal linear networks, the linear stability is critical in characterizing the sharpness of minima found by SGD, which is consistent with the fact that in this case, SGD converges to global minima stably. As a comparison, the right panel of Figure 3(b) shows that for GD, all the three sharpness keep almost unchanged when increasing the LR.

The test performance. Figure 3(b) shows the test errors of minima found by SGD and GD for varying LRs. One can see that as increasing the LR, the test error of SGD decreases significantly. This can be explained by the fact that $\mathrm{Tr}(G(\theta))$ and the resulting ℓ_1 norm of β decrease significantly as demonstrated in Figure 3(a). In contrast, the test error of GD barely changes, which is also consistent with our theoretical prediction that the stability of GD can not yield effective capacity control for diagonal linear networks. These are consistent with our theoretical prediction: The stability-induced regularization of SGD is much stronger than that of GD.

The balancing effect. To measure the balancedness between the inner and outer layers, we define $r(\theta) = \|\alpha\|_1/2\|\beta\|_1 = \sum_j (a_j^2 + b_j^2)/(2\sum_j |a_j b_j|)$. By the AM-GM inequality, $r(\theta) \geqslant 1$ and the equality is reached when $a_j^2 = b_j^2$ for all $j \in [n]$, i.e., the solutions are totally balanced. The larger $r(\theta)$ is, the less balanced the solution is. In the experiment, we consider the initialization $a_j \sim \mathcal{N}(0, 0.1), b_j \sim \mathcal{N}(0, 0.1 r_0)$ with r_0 controlling the balancedness at initialization. We are interested in the balancedness of minima selected by SGD and GD. Figure 3(c) shows that SGD finds solutions with $r(\theta) \approx 1$ no matter how unbalanced the initialization is. In contrast, GD is expectedly unable to reduce the unbalancedness introduced at initialization. These confirm again our theoretical predictions by analyzing the dynamical stability.

6. Conclusion

In this paper, we study the stability-induced regularization of SGD and GD by relating the dynamical stability to the sharpness of local landscape. We establish generalization bounds of stable minima for two-layer ReLU networks and diagonal linear networks via linking sharpness to parameter norms. Specifically, these bounds imply that stable minima of SGD provably generalize well and can explain the benefit of using a large LR. Most importantly, our stability analysis can explain why SGD generalizes better than GD at least for diagonal linear networks. We also corroborate our theoretical findings with fine-grained numerical experiments.

Note that the stability-induced regularization is independent of initialization but crucially depends on the size of LR. This can potentially explain the practical observation that large LR often leads to better generalization in training large-scale models. In contrast, other mechanisms such as small initialization (Chizat & Bach, 2020; Woodworth et al., 2020) and noise-driven diffusion (Blanc et al., 2020; Li et al., 2021; Damian et al., 2021) cannot explain the benefit of large LR. In addition, our analysis also suggests that gradient clipping has an implicit regularization effect in the way of allowing convergence with a larger LR. We leave the systematic investigation of these issues to future work.

Acknowledgements

The work of Lei Wu is supported by a startup fund from Peking University. The work of Weijie J. Su is supported in part by NSF Grants CAREER DMS-1847415 and an Alfred Sloan Research Fellowship. We thank Yaroslav Bulatov for bringing the reference (Défossez & Bach, 2015) to our attention and many helpful discussions. We also thank the anonymous reviewers for their valuable suggestions.

References

Barrett, D. and Dherin, B. Implicit gradient regularization. In *International Conference on Learning Representations*, 2020. 3

Barron, A. R. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, 39(3):930–945, 1993. 2

Bartlett, P. L., Bousquet, O., and Mendelson, S. Local Rademacher complexities. *The Annals of Statistics*, 33 (4):1497–1537, 2005. 13

Blanc, G., Gupta, N., Valiant, G., and Valiant, P. Implicit regularization for deep neural networks driven by an Ornstein-Uhlenbeck like process. In *Conference on learning theory*, pp. 483–513. PMLR, 2020. 2, 6, 9

Chizat, L. and Bach, F. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on Learning Theory*, pp. 1305–1338. PMLR, 2020. 2, 9

Damian, A., Ma, T., and Lee, J. Label noise SGD provably prefers flat global minimizers. *arXiv preprint arXiv:2106.06530*, 2021. 9

Défossez, A. and Bach, F. Averaged least-mean-squares: Bias-variance trade-offs and optimal sampling distributions. In *Artificial Intelligence and Statistics*, pp. 205–213. PMLR, 2015. 4, 9

Dinh, L., Pascanu, R., Bengio, S., and Bengio, Y. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, pp. 1019–1028. PMLR, 2017. 3

E, W., Ma, C., and Wu, L. A priori estimates of the population risk for two-layer neural networks. *Communications in Mathematical Sciences*, 17(5):1407–1425, 2019. 2

E, W., Ma, C., and Wu, L. The Barron space and the flow-induced function spaces for neural network models. *Constructive Approximation*, pp. 1–38, 2021. 5

Feng, Y. and Tu, Y. The inverse variance–flatness relation in stochastic gradient descent is critical for finding flat minima. *Proceedings of the National Academy of Sciences*, 118(9), 2021. 4

Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2020. 1

Gissin, D., Shalev-Shwartz, S., and Daniely, A. The implicit bias of depth: How incremental learning drives generalization. In *International Conference on Learning Representations*, 2019. 7

He, H. and Su, W. The local elasticity of neural networks. In *International Conference on Learning Representations*, 2020. 2

Hochreiter, S. and Schmidhuber, J. Simplifying neural nets by discovering flat minima. *Advances in neural information processing systems*, 7, 1994. 1

Hochreiter, S. and Schmidhuber, J. Flat minima. *Neural Computation*, 9(1):1–42, 1997. 1

Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., and Wilson, A. G. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018. 1

Jastrzębski, S., Kenton, Z., Arpit, D., Ballas, N., Fischer, A., Bengio, Y., and Storkey, A. Three factors influencing minima in SGD. *arXiv preprint arXiv:1711.04623*, 2017. 1

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. On large-batch training for deep learning: Generalization gap and sharp minima. In *In International Conference on Learning Representations (ICLR)*, 2017. 1

Li, Z., Wang, T., and Arora, S. What happens after SGD reaches zero loss?–a mathematical framework. *arXiv preprint arXiv:2110.06914*, 2021. 9

Liang, T., Poggio, T., Rakhlin, A., and Stokes, J. Fisher-Rao metric, geometry, and complexity of neural networks. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89, pp. 888–896. PMLR, 2019. 3

Liu, K., Ziyin, L., and Ueda, M. Noise and fluctuation of finite learning rate stochastic gradient descent. In *International Conference on Machine Learning*, pp. 7045–7056. PMLR, 2021. 4

Ma, C. and Ying, L. On linear stability of SGD and input-smoothness of neural networks. *Advances in Neural Information Processing Systems*, 34, 2021. 2, 3, 4, 6

Ma, C., Wu, L., and E, W. The quenching-activation behavior of the gradient descent dynamics for two-layer neural network models. *arXiv preprint arXiv:2006.14450*, 2020. 2

McAllester, D. A. Some PAC-Bayesian theorems. *Machine Learning*, 37(3):355–363, 1999. 3

Mikolov, T. et al. Statistical language models based on neural networks. *Presentation at Google, Mountain View, 2nd April*, 80(26), 2012. 3

Mori, T., Ziyin, L., Liu, K., and Ueda, M. Power-law escape rate of SGD. In *International Conference on Machine Learning*, pp. 15959–15975. PMLR, 2022. 4

Mulayoff, R., Michaeli, T., and Soudry, D. The implicit bias of minima stability: A view from function space. *Advances in Neural Information Processing Systems*, 34, 2021. 2, 3, 6

Nacson, M. S., Ravichandran, K., Srebro, N., and Soudry, D. Implicit bias of the step size in linear diagonal neural networks. In *International Conference on Machine Learning*, pp. 16270–16295. PMLR, 2022. 2, 3, 7, 8

Neyshabur, B., Tomioka, R., and Srebro, N. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014. 1

Neyshabur, B., Tomioka, R., and Srebro, N. Norm-based capacity control in neural networks. In *Conference on Learning Theory*, pp. 1376–1401. PMLR, 2015. 2, 5

Neyshabur, B., Bhojanapalli, S., Mcallester, D., and Srebro, N. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems 30*, pp. 5949–5958, 2017. 3

O'Donnell, R. *Analysis of boolean functions*. Cambridge University Press, 2014. 12

Pascanu, R., Mikolov, T., and Bengio, Y. On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pp. 1310–1318. PMLR, 2013. 3

Pesme, S., Pillaud-Vivien, L., and Flammarion, N. Implicit bias of SGD for diagonal linear networks: a provable benefit of stochasticity. *arXiv preprint arXiv:2106.09524*, 2021. 2, 7

Petzka, H., Kamp, M., Adilova, L., Sminchisescu, C., and Boley, M. Relative flatness and generalization. *Advances in Neural Information Processing Systems*, 34, 2021. 3

Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014. 12, 28

Smith, S. L., Dherin, B., Barrett, D., and De, S. On the origin of implicit regularization in stochastic gradient descent. In *International Conference on Learning Representations*, 2020. 3

Srebro, N., Sridharan, K., and Tewari, A. Smoothness, low noise and fast rates. *Advances in neural information processing systems*, 23, 2010. 13

Su, W. Neurashed: A phenomenological model for imitating deep learning training. *arXiv preprint arXiv:2112.09741*, 2021. 2

Tsuzuku, Y., Sato, I., and Sugiyama, M. Normalized flat minima: Exploring scale invariant definition of flat minima for neural networks using PAC-Bayesian analysis. In *International Conference on Machine Learning*, pp. 9636–9647. PMLR, 2020. 3

Vardi, G. On the implicit bias in deep-learning algorithms. *arXiv preprint arXiv:2208.12591*, 2022. 2

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018. 14, 15

Wojtowytsch, S. Stochastic gradient descent with noise of machine learning type. part II: Continuous time analysis. *arXiv preprint arXiv:2106.02588*, 2021. 4

Woodworth, B., Gunasekar, S., Lee, J. D., Moroshko, E., Savarese, P., Golan, I., Soudry, D., and Srebro, N. Kernel

and rich regimes in overparametrized models. In *Conference on Learning Theory*, pp. 3635–3673. PMLR, 2020. 2, 7, 9

Wu, D., Xia, S.-T., and Wang, Y. Adversarial weight perturbation helps robust generalization. *Advances in Neural Information Processing Systems*, 33, 2020a. 1

Wu, J., Hu, W., Xiong, H., Huan, J., Braverman, V., and Zhu, Z. On the noisy gradient descent that generalizes as SGD. In *International Conference on Machine Learning*, pp. 10367–10376. PMLR, 2020b. 1

Wu, L., Zhu, Z., and E, W. Towards understanding generalization of deep learning: Perspective of loss landscapes. *arXiv preprint arXiv:1706.10239*, 2017. 1, 2

Wu, L., Ma, C., and E, W. How SGD selects the global minima in over-parameterized learning: A dynamical stability perspective. *Advances in Neural Information Processing Systems*, 31:8279–8288, 2018. 1, 2, 3

Wu, L., Wang, M., and Su, W. J. The alignment property of SGD noise and how it helps select flat minima: A stability analysis. In *Advances in Neural Information Processing Systems*, 2022. 1, 2, 4, 5

Xu, Z.-Q. J., Zhou, H., Luo, T., and Zhang, Y. Towards understanding the condensation of two-layer neural networks at initial training. *arXiv preprint arXiv:2105.11686*, 2021. 2

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017. 1, 2

Zhu, Z., Wu, J., Yu, B., Wu, L., and Ma, J. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects. In *International Conference on Machine Learning*, pp. 7654–7663. PMLR, 2019. 5

Appendix

A. Technical background

In this section, we will first introduce some notations and technical background which will be used in the proofs of next sections.

A.1. The Hermite expansion.

Let $\gamma = \mathcal{N}(0,1)$ and $\{h_i\}_{i=0}^{\infty}$ be the probabilist's Hermite polynomials, which form a set of orthonormal basis of $L^2(\gamma)$ with

$$h_0(z) = 1,\ h_1(z) = z,\ h_2(z) = \frac{z^2-1}{\sqrt{2}}, h_3(z) = \frac{z^3-3z}{\sqrt{6}}, \cdots . \tag{13}$$

Given a $f \in L^2(\gamma)$, denote by $f(z) = \sum_k \hat{f}_k h_k(z)$ be the Hermite expansion of f where

$$\hat{f}_k = \mathbb{E}_{z\sim\gamma}[f(z)h_k(z)] = \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} f(z)h_k(z)\,\mathrm{d}z$$

is the "Fourier coefficient" of f. We will frequently use the following lemma (O'Donnell, 2014, Proposition 11.31):

Lemma A.1. *Given $f, g \in L^2(\gamma)$, we have for any $u, v \in \mathbb{S}^{d-1}$ that*

$$\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}[f(u^Tx)g(v^Tx)] = \sum_{k=0}^{\infty} \hat{f}_k\hat{g}_k(u^Tv)^k.$$

A.2. Rademacher complexity and generalization bounds

Here we only state properties of Rademacher complexity that will be used in this paper. For the missing proofs and more details, we refer to Shalev-Shwartz & Ben-David (2014, Section 26).

Definition A.2. Given a function class $\mathcal{F}$, the Rademacher complexity of $\mathcal{F}$ with respect to $x_1, \ldots, x_n$ is defined as

$$\widehat{\mathrm{Rad}}_n(\mathcal{F}) = \mathbb{E}_{\xi_1,\ldots,\xi_n}[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} f(x_i)\xi_i],$$

where $\xi_1, \ldots, \xi_n$ are i.i.d. samples drawn from the Rademacher distribution: $\mathbb{P}(\xi = 1) = \mathbb{P}(\xi = -1) = \frac{1}{2}$.

Lemma A.3 (Contraction property)**.** *Let $\varphi : \mathbb{R} \mapsto \mathbb{R}$ be β-Lispchitz continuous and $\varphi \circ \mathcal{F} = \{\varphi \circ f : f \in \mathcal{F}\}$. Then, $\widehat{\mathrm{Rad}}_n(\varphi \circ \mathcal{F}) \leqslant \beta\,\widehat{\mathrm{Rad}}_n(\mathcal{F})$.*

Lemma A.4. *Let $\mathcal{F} = \{u^Tx : u \in \mathbb{S}^{d-1}\}$ be the linear class. Then $\widehat{\mathrm{Rad}}_n(\mathcal{F}) \leqslant \sqrt{\frac{\sum_{i=1}^n \|x_i\|^2}{n^2}}$.*

Theorem A.5. *Consider a function class $\mathcal{F}$ with $\sup_{z\in\mathcal{X}, f\in\mathcal{F}} |f(z)| \leqslant B$. For any $\delta \in (0,1)$, w.p. at least $1-\delta$ over the choice of $S = (z_1, z_2, \ldots, z_n)$, we have,*

$$|\frac{1}{n}\sum_{i=1}^{n} f(z_i) - \mathbb{E}_z[f(z)]| \lesssim \widehat{\mathrm{Rad}}_n(\mathcal{F}) + B\sqrt{\frac{\ln(2/\delta)}{n}}.$$

Lemma A.6. *Let $\mathcal{F}$ and $\mathcal{G}$ be two function classes. Suppose that $\sup_{f\in\mathcal{F}} \|f\|_\infty \leqslant A$ and $\sup_{g\in\mathcal{G}} \|g\|_\infty \leqslant B$. Define $\mathcal{F} * \mathcal{G} = \{f(x)g(x) : \mathcal{X} \mapsto \mathbb{R} :\ f \in \mathcal{F}, g \in \mathcal{G}\}$. Then, $\widehat{\mathrm{Rad}}_n(\mathcal{F} * \mathcal{G}) \leqslant (A+B)(\widehat{\mathrm{Rad}}_n(\mathcal{F}) + \widehat{\mathrm{Rad}}_n(\mathcal{G}))$.*

Proof. By the definition of Rademacher complexity,

$$\begin{aligned}
n\widehat{\mathrm{Rad}}_n(\mathcal{F}*\mathcal{G}) &= \mathbb{E}_\xi[\sup_{f\in\mathcal{F},g\in\mathcal{G}}\sum_{i=1}^n f(x_i)g(x_i)\xi_i] \\
&= \mathbb{E}_\xi[\sup_{f\in\mathcal{F},g\in\mathcal{G}}\sum_{i=1}^n \frac{(f(x_i)+g(x_i))^2}{4}\xi_i - \sum_{i=1}^n \frac{(f(x_i)-g(x_i))^2}{4}\xi_i] \\
&\leqslant \mathbb{E}_\xi[\sup_{f\in\mathcal{F},g\in\mathcal{G}}\sum_{i=1}^n \frac{(f(x_i)+g(x_i))^2}{4}\xi_i + \mathbb{E}_\xi[\sup_{f\in\mathcal{F},g\in\mathcal{G}}\sum_{i=1}^n \frac{(f(x_i)-g(x_i))^2}{4}\xi_i] \\
&\overset{(i)}{\leqslant} \frac{A+B}{2}\left(\mathbb{E}_\xi[\sup_{f\in\mathcal{F},g\in\mathcal{G}}\sum_{i=1}^n (f(x_i)+g(x_i))\xi_i] + \mathbb{E}_\xi[\sup_{f\in\mathcal{F},g\in\mathcal{G}}\sum_{i=1}^n (f(x_i)-g(x_i))\xi_i]\right) \\
&\leqslant (A+B)n(\widehat{\mathrm{Rad}}_n(\mathcal{F}) + \widehat{\mathrm{Rad}}_n(\mathcal{G})),
\end{aligned}$$

where (i) follows from the Lemma A.3 and the fact that $t^2/4$ is $(A+B)/2$ Lipschitz continuous since $|f|\leqslant A, |g|\leqslant B$. □

Generalization bounds of learning with a smooth loss. Let $\phi : \mathcal{Y}\times\mathcal{Y}\mapsto[0,\infty)$ be a loss function. Define the empirical and population risk as follows

$$\hat{\mathcal{R}}(h) = \hat{\mathbb{E}}[\phi(h(x),y)], \quad \mathcal{R}(h) = \mathbb{E}[\phi(h(x),y)],$$

where $\hat{\mathbb{E}}$ denotes the expectation with respect to the empirical measure. Let $\mathcal{H}$ be the hypothesis space and $\hat{h} = \mathrm{argmin}_{h\in\mathcal{H}}\hat{L}(h)$. We would like to bound the population risk of the $\hat{h}$ by using the following decomposition:

$$\mathcal{R}(\hat{h}) = \hat{\mathcal{R}}(\hat{h}) + \underbrace{\mathcal{R}(\hat{h}) - \hat{\mathcal{R}}(\hat{h})}_{gen-gap}.$$

Theorem A.5 shows that the second term (gen-gap) can be controlled by the Rademacher complexity of $\mathcal{H}$. By assuming ϕ is Lipschitz continuous and applying Lemma A.3, an (informal) bound goes like

$$\mathcal{R}(\hat{h}) - \hat{\mathcal{R}}(\hat{h}) \leqslant \sup_{h\in\mathcal{H}}|\mathcal{R}(h) - \hat{\mathcal{R}}(h)| \leqslant Lip(\phi)\widehat{\mathrm{Rad}}_n(\mathcal{H}).$$

This usually provides us a $O(1/\sqrt{n})$ bound, which is tight for Lipschitz loss such as hinge loss. However, for square loss, this bound is often loose as explained as follows. For the minimizer $\hat{h}$, it is expected that $\mathcal{R}(\hat{h})\leqslant r$ for small r. Therefore, one only needs to consider a constraint hypothesis class:

$$\mathcal{H}_r = \{h\in\mathcal{H}\,|\,\mathcal{R}(h)\leqslant r\}.$$

For hypothesis in this restricted class, the Lipschitz constant of ϕ is much smaller for smooth loss. For instance, $t^2/2$ is only r-Lipschitz for $t\in[-r,r]$. This argument can be formalized by using the concept of local Rademacher complexity (Bartlett et al., 2005). Specifically, we shall use the following theorem in our proof, which is a restatement of Srebro et al. (2010, Theorem 1)

Theorem A.7. *Let*

$$\mathfrak{R}_n(\mathcal{H}) = \sup_{x_1,\ldots,x_n}\widehat{\mathrm{Rad}}_n(\mathcal{H}) \tag{14}$$

be the worst-case Rademacher complexity. Assume that $|\phi''|\leqslant A$ and $0\leqslant\phi\leqslant B$.. Then, w.p. at least $1-\delta$ over the sampling of training set, we have for any $h\in\mathcal{H}$ that

$$\mathcal{R}(h)\leqslant\hat{\mathcal{R}}(h) + C\left(\sqrt{\hat{\mathcal{R}}(h)}\left(\sqrt{A}\log^{3/2}(n)\mathfrak{R}_n(\mathcal{H}) + \sqrt{\frac{B\log(1/\delta)}{n}}\right) + A\log^3(n)\mathfrak{R}_n(\mathcal{H})^2 + \frac{B\log(1/\delta)}{n}\right).$$

In particular, for $\hat{h}\in\mathrm{argmin}_h\hat{\mathcal{R}}(h)$,

$$\mathcal{R}(\hat{h})\lesssim A\log^3(n)\mathfrak{R}_n(\mathcal{H})^2 + \frac{B\log(1/\delta)}{n}. \tag{15}$$

In this paper, we will mainly use (15) to bound generalization error since our focus is the minimizer $\hat{h}$.

A.3. Concentration inequalities

Definition A.8. Let ψ be a non-decreasing, convex function with $\psi(0) = 0$. The Orlicz norm of a random variable X is defined by $\|X\|_\psi := \inf\{t > 0 : \mathbb{E}[\psi(|X|/t)] \leqslant 1\}$. If $X \in \mathbb{R}^d$ is a vector, then $\|X\|_\psi := \sup_{u\in\mathbb{S}^{d-1}} \|u^T X\|_\psi$.

For our purpose, Orlicz norms of interest are the ones given by $\psi_p(x) = e^{x^p} - 1$ for $p \geqslant 1$. In particular, the cases of $p = 1$ and $p = 2$ correspond to the sub-exponential and sub-Gaussian norms, respectively. A random variable X is said to be sub-Gaussian (resp. sub-exponential) if $\|X\|_{\psi_2} < \infty$ (resp. $\|X\|_{\psi_1} < 1$).

A random variable with finite ψ_p-norm has the following control of the tail behavior

$$\mathbb{P}\{|X| \geqslant t\} \leqslant C_1 e^{-C_2 \frac{t^p}{\|X\|_{\psi_p}^p}},$$

where C_1, C_2 are constant that only depend on p.

Lemma A.9.

- *If $|X| \lesssim 1$ almost surely, then $\|X\|_{\psi_i} \lesssim 1$ for $i = 1, 2$.*
- *If $X \sim \mathcal{N}(0, \sigma^2)$, X is sub-Gaussian with $\|X\|_{\psi_2} \leqslant C\sigma$.*
- *Let X, Y be sub-Gaussian random variables. Then, XY is sub-exponential and $\|XY\|_{\psi_1} \leqslant \|X\|_{\psi_2}\|Y\|_{\psi_2}$.*
- *If $|X| \leqslant |Y|$ a.s., then $\|X\|_\psi \leqslant \|Y\|_\psi$ for any ψ that satisfies the condition in Definition A.8.*
- ***Center inequality.** For a random variable X, we have*

$$\|X - \mathbb{E}[X]\|_{\psi_p} \leqslant C\|X\|_{\psi_p} \tag{16}$$

for a constant $C > 0$ that may depend on p.

Theorem A.10 (Bernstein's inequality)**.** *Let $X_1, \ldots, X_n$ be independent sub-exponential random variables. Suppose $K = \max_i \|X_i\|_{\psi_1} < \infty$. Then, for any $t > 0$,*

$$\mathbb{P}\Big\{\big|\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}[X]\big| \geqslant t\Big\} \leqslant 2\exp\left(-Cn\min\left(\frac{t^2}{K^2}, \frac{t}{K}\right)\right).$$

Proposition A.11 (Sums of independent sub-Gaussians)**.** *Let $X_1, \ldots, X_n$ be independent, mean zero, sub-Gaussian random variables. Then, $\sum_{i=1}^n X_i$ is also a sub-Gaussian random variable, and*

$$\|\sum_{i=1}^n X_i\|_{\psi_2}^2 \leqslant C\sum_{i=1}^n \|X_i\|_{\psi_2}^2.$$

Covering number. We shall also use the covering number in our analysis. Let (T, q) be a metric space. Consider a subset $K \subset T$ and let $\varepsilon > 0$. A subset $\mathcal{N}_\varepsilon$ is called an ε-net of K if every point in K is within a distance ε of some point of $\mathcal{N}_\varepsilon$, i.e.,

$$\forall x \in K,\, \exists\, x_0 \in \mathcal{N}_\varepsilon \,:\, q(x, x_0) \leqslant \varepsilon.$$

The smallest possible cardinality of an ε-net of K is called the covering number of K and is denoted by $N(K, q, \varepsilon)$.

A commonly-used fact is

$$N(\mathbb{S}^{d-1}, \|\cdot\|, \varepsilon) \leqslant (1 + 2/\varepsilon)^d \tag{17}$$

(see, e.g., Vershynin (2018, Corollary 4.2.13)).

Remark: We refer the reader to Vershynin (2018) for the proofs of the above properties and more related information.

A.4. Auxiliary Lemmas

Lemma A.12. *Let $u_1, u_2, \dots, u_m \in \mathbb{R}^d$. Then for any $k \in \mathbb{N}$ and $\alpha \in \mathbb{R}^m$, we have*

$$\sum_{i,j=1}^{m} \alpha_i \alpha_j (u_i^T u_j)^k \geqslant 0. \tag{18}$$

Proof. Let $U = (u_1, \dots, u_m) \in \mathbb{R}^{d\times m}$ and $Q_k = ((u_i^T u_j)^k)_{i,j} \in \mathbb{R}^{m\times m}$. First, $Q_0 = I_d, Q_1 = U^T U$ are both positive semi-definite and hence (18) holds. For $k \geqslant 2$, we have $Q_k = Q_1 \circ Q_1 \circ \cdots \circ Q_1$ where $\circ$ denotes the hadamard product. By the Schur product theorem[1], Q_k is also positive semi-definite and hence (18) holds. □

Lemma A.13. *Suppose $k(\cdot,\cdot)$ to be positive semi-definite kernel and let $\phi : \mathcal{X} \mapsto \mathcal{H}$ be a feature map satisfying $k(x,y) = \langle \phi(x), \phi(y)\rangle_{\mathcal{H}}$. Then,*

$$\lambda_1(\mathcal{K}) = \sup_{\|h\|_{\mathcal{H}}=1} \mathbb{E}_x[\langle h, \phi(x)\rangle_{\mathcal{H}}^2]. \tag{19}$$

Proof. By the variational principle of the largest eigenvalue, we have

$$\begin{aligned}
\lambda_1(\mathcal{K}) &= \sup_{\|u\|_{L_2(\rho)}=1} \mathbb{E}_{x,y}[k(x,y)u(x)u(y)] = \sup_{\|u\|_{L_2(\rho)}=1} \mathbb{E}_{x,y}[\langle \phi(x), \phi(y)\rangle_{\mathcal{H}} u(x)u(y)] \\
&= \sup_{\|u\|_{L_2(\rho)}=1} \|\mathbb{E}_x[u(x)\phi(x)]\|_{\mathcal{H}}^2 = \sup_{\|u\|_{L_2(\rho)}=1} \sup_{\|h\|_{\mathcal{H}}=1} \langle h, \mathbb{E}_x[u(x)\phi(x)]\rangle_{\mathcal{H}}^2 \\
&= \sup_{\|h\|_{\mathcal{H}}=1} \sup_{\|u\|_{L_2(\rho)}=1} \mathbb{E}_x[u(x)\langle h, \phi(x)\rangle_{\mathcal{H}}]^2 = \sup_{\|h\|_{\mathcal{H}}=1} \mathbb{E}_x[\langle h, \phi(x)\rangle_{\mathcal{H}}^2].
\end{aligned}$$

□

Lemma A.14. *Assume $X \sim \mathcal{N}(0, I_d)$. For any $\delta \in (0,1)$, let $n \gtrsim d + \log(1/\delta)$, then w.p. $1-\delta$, we have*

$$\|\hat{\Sigma}_n - \Sigma\|_2 \lesssim \sqrt{\frac{d+\log(1/\delta)}{n}} + \frac{d+\log(1/\delta)}{n}, \qquad \|\hat{\Sigma}_n\|_2 \lesssim 1, \qquad \|\hat{\Sigma}_n\|_F \lesssim \sqrt{d}.$$

Proof. First by Vershynin (2018, Exercise 4.7.4), for any $\delta \in (0,1)$, w.p. $1-\delta$ it holds that

$$\|\hat{\Sigma}_n - \Sigma\|_2 \lesssim \left(\sqrt{\frac{d+\log(1/\delta)}{n}} + \frac{d+\log(1/\delta)}{n}\right) \|\Sigma\|_2.$$

Therefore, if $n \gtrsim d + \log(1/\delta)$, we have

$$\|\hat{\Sigma}_n\|_2 \leqslant \|\hat{\Sigma}_n - \Sigma\|_2 + \|\Sigma\|_2 \lesssim \|\Sigma\|_2 = 1.$$

Moreover,

$$\|\hat{\Sigma}_n\|_F^2 = \sum_{j=1}^{d} \lambda_j^2(\hat{\Sigma}_n) \leqslant d\|\hat{\Sigma}_n\|_2 \lesssim d.$$

Taking the square root completes the proof.

□

[1] see https://en.wikipedia.org/wiki/Schur_product_theorem

B. Missing Proofs of Section 3

B.1. Proof of Proposition 3.2.

Let $\mathcal{S}_+$ be the set of $p \times p$ positive semi-definite matrices, $H_i = g_i(\theta^*)g_i(\theta^*)^T$ for $i \in [n]$ and $\delta_t = \theta_t - \theta^*$. The linearized SGD (4) can be rewritten as $\delta_{t+1} = (I - \eta H_{i_t})\delta_t$ with $i_t \in \mathrm{Unif}([n])$. Then,

$$\begin{aligned}\delta_{t+1}\delta_{t+1}^T &= (\delta_t - \eta H_{i_t}\delta_t)(\delta_t - \eta H_{i_t}\delta_t)^T \\ &= \delta_t\delta_t^T - \eta(\delta_t\delta_t^T H_{i_t} + H_{i_t}\delta_t\delta_t^T) + \eta^2 H_{i_t}\delta_t\delta_t^T H_{i_t}.\end{aligned}$$

Let $Q_t = \mathbb{E}[\delta_t\delta_t^T]$ be the deviation covariance matrix. Then taking expectation gives

$$\begin{aligned}Q_{t+1} &= Q_t - \eta(Q_t H + HQ_t) + \eta^2\,\mathbb{E}[H_{i_t}Q_t H_{i_t}] \\ &= (I - \eta T_\eta)Q_t,\end{aligned} \tag{20}$$

where $T_\eta : \mathcal{S}_+ \mapsto \mathcal{S}_+$ is given by $T_\eta A = (HA + AH) - \eta\,\mathbb{E}[H_\xi A H_\xi]$, where the expectation is taken with respect to $\xi \sim \mathrm{Unif}([n])$.

By (20), to ensure $\|\mathbb{E}[Q_t]\|_F \leqslant C\|\mathbb{E}[Q_0]\|_F$ for some constant $C > 0$, we need $T_\eta \succeq 0$. This is equivalent to it holds that

$$\langle A, T_\eta A\rangle = 2\mathrm{Tr}(AHA) - \eta\,\mathbb{E}[\mathrm{Tr}(AH_\xi)AH_\xi] \geqslant 0 \quad \forall\, A \in \mathcal{S}_+. \tag{21}$$

Noticing that $\mathbb{E}[\mathrm{Tr}(AH_\xi AH_\xi)] = \mathbb{E}[(g_\xi^T A g_\xi)^2] \geqslant (\mathbb{E}[g_\xi A g_\xi^T])^2 = \mathrm{Tr}^2(HA)$, (21) implies

$$2\mathrm{Tr}(HA^2) - \eta\mathrm{Tr}^2(AH) \geqslant 0, \quad \forall A \in \mathbb{S}_+.$$

Taking $A = \mathrm{diag}(w_1, \ldots, w_p)$, we obtain

$$\frac{(\sum_j \lambda_j(H)w_j)^2}{\sum_j \lambda_j(H)w_j^2} \leqslant \frac{2}{\eta}. \tag{22}$$

Specifically, taking $w_j = 1$ for $j = 1, \ldots, n$ completes the proof. □

Remark B.1. It should be stressed that the stability condition (22) is stronger than $\mathrm{Tr}(H) \leqslant 2/\eta$. We only state the latter in the main text since it is more intuitive and has clean relationship to sharpness.

B.2. Proof of Proposition 3.4

By (2), we have $H(\theta) = G(\theta) + O(\varepsilon^{1/2})$ for any $\theta \in \mathcal{Q}_{\varepsilon,\eta}$. Then, we have

$$\mathrm{Tr}[H(\theta_t)S(\theta_t)] = \mathrm{Tr}[G(\theta_t)S(\theta_t)] + O(\varepsilon^{1/2}) \geqslant \mu_0\|G(\theta_t)\|_F^2 + O(\varepsilon^{1/2}), \tag{23}$$

where the second step follows from the definition of $\mu(\theta)$ and the assumption that $\mu(\theta) \geqslant \mu_0$.

Let $\gamma := \eta^2 \inf_{\theta\in\mathcal{Q}_{\varepsilon,\eta}} \mu_0\|G(\theta)\|_F^2$. Then combining Lemma 3.3 and (23) gives

$$\begin{aligned}\mathbb{E}[\hat{\mathcal{R}}(\theta_t)] &\geqslant \eta^2\mu_0\|G(\theta_t)\|_F^2\,\mathbb{E}[\hat{\mathcal{R}}(\theta_t)] + O(\eta^3 + \eta^2\varepsilon^{3/2}) \\ &\geqslant \gamma\,\mathbb{E}[\hat{\mathcal{R}}(\theta_t)] + O(\eta^3 + \eta^2\varepsilon^{3/2}) \\ &\geqslant \gamma^t\,\mathbb{E}[\hat{\mathcal{R}}(\theta_0)] + \frac{\gamma^t - 1}{\gamma - 1}O(\eta^3 + \eta^2\varepsilon^{3/2}).\end{aligned}$$

□

C. Missing Proofs in Section 4

In this section, we will frequently use the following definition and results.

- **Kernel functions.** Define two associated kernel functions:

$$\varphi_1(u, v) := \mathbb{E}_x[\sigma(u^T x)\sigma(v^T x)], \quad \varphi_2(u, v) := \mathbb{E}_x[\sigma'(u^T x)\sigma'(v^T x)], \tag{24}$$

where $\varphi_1, \varphi_2 : \Omega \mapsto \mathbb{R}$ with $\Omega := \mathbb{S}^{d-1} \otimes \mathbb{S}^{d-1}$. The corresponding empirical ones are given by

$$\hat{\varphi}_1(u,v) = \frac{1}{n}\sum_{i=1}^{n} \sigma(u^T x_i)\sigma(v^T x_i), \quad \hat{\varphi}_2(u,v) = \frac{1}{n}\sum_{i=1}^{n} \sigma'(u^T x_i)\sigma'(v^T x_i). \tag{25}$$

- **Hermite expansions of kernels.** Let $\sigma(t) = \sum_{k=0}^{\infty} \alpha_k h_k(t)$ and $\sigma'(t) = \sum_k \beta_k h_k(t)$ be the Hermite expansions of σ and σ', respectively.

Lemma C.1. *Define* $\|(u,v) - (u',v')\|_\Omega = \|u-u'\| + \|v-v'\|$ *for any* $(u,v),(u',v') \in \Omega$. *Then,*

$$\mathcal{N}(\Omega, \|\cdot\|_\Omega, \epsilon) \leqslant \mathcal{N}(\mathbb{S}^{d-1}, \|\cdot\|, \epsilon/2)^2 \leqslant (6/\epsilon)^{2d}. \tag{26}$$

Proof. Follow trivially from the fact (17). □

Lemma C.2 (Property of kernel functions)**.**

- $\varphi_1(u,v) = \|u\|\|v\| \sum_k \alpha_k^2 (\hat{u}^T\hat{v})^k, \varphi_2(u,v) = \sum_k \beta_k^2 (\hat{u}^T\hat{v})^k$.
- $\alpha_0 = \beta_1 \sim 1$ *and* $\beta_0 \sim 1$.
- $\varphi_i(u,v) \sim 1$ *for any* $u, v \in \mathbb{S}^{d-1}$ *and* $i = 1, 2$.
- *For any* $s \in \mathbb{R}^m$ *and* $u_1, u_2, \ldots, u_m \in \mathbb{S}^{d-1}$*, we have*

$$\sum_{j,k=1}^{m} s_j^2 s_k^2 \varphi_i(u_j^T u_k) \gtrsim (\sum_{j=1}^{m} s_j^2)^2, \quad \forall i = 1, 2.$$

Proof.

- Using the positive homogeneity of σ and Lemma A.1, we have

$$\begin{aligned}\varphi_1(u,v) &= \|u\|\|v\| \, \mathbb{E}_x[\sigma(\hat{u}^T x)\sigma(\hat{v}^T x)] = \|u\|\|v\| \, \mathbb{E}_x[\sum_k \alpha_k h_k(\hat{u}^T x) \sum_l \alpha_l h_l(\hat{v}^T x)] \\ &= \|u\|\|v\| \sum_{k,l} \alpha_k \alpha_l \delta_{k,l} (\hat{u}^T\hat{v})^k = \|u\|\|v\| \sum_k \alpha_k^2 (\hat{u}^T\hat{v})^k. \end{aligned} \tag{27}$$

  Similarly, we have the expansion of φ_2.

- Noticing $h_0(z) = 1$ and $h_1(z) = z$, we have

$$\begin{aligned}\alpha_0 &= \frac{1}{\sqrt{2\pi}} \int \sigma(t) e^{-t^2/2} \, \mathrm{d}t \sim \int_0^\infty t e^{-t^2/2} \, \mathrm{d}t \sim 1 \\ \beta_0 &= \frac{1}{\sqrt{2\pi}} \int \sigma'(t) e^{-t^2/2} \, \mathrm{d}t \sim \frac{1}{\sqrt{2\pi}} \int_0^\infty e^{-t^2/2} \, \mathrm{d}t \sim 1 \\ \beta_1 &= \frac{1}{\sqrt{2\pi}} \int \sigma'(t) t e^{-t^2/2} \, \mathrm{d}t = \frac{1}{\sqrt{2\pi}} \int \sigma(t) e^{-t^2/2} \, \mathrm{d}t = \alpha_0 \sim 1. \end{aligned}$$

- We now prove the third conclusion. By (27), we have

$$\begin{aligned}\sum_{j,k=1}^{m} s_j^2 s_k^2 \varphi_1(u_j^T u_k) &= \sum_{j,k=1}^{m} s_j^2 s_k^2 \sum_{l=0}^{\infty} \alpha_l^2 (u_j^T u_k)^l = \alpha_0^2 \sum_{j,k=1}^{m} s_j^2 s_k^2 + \sum_{l=1}^{\infty} \alpha_l^2 \sum_{j,k=1}^{m} s_j^2 s_k^2 (u_j^T u_k)^l \\ &\geqslant \alpha_0^2 \sum_{j,k=1}^{m} s_j^2 s_k^2 + 0 \qquad \text{(use Lemma A.12)} \\ &\gtrsim (\sum_j s_j^2)^2, \end{aligned}$$

  where the last step is due to $\alpha_0 \sim 1$. The case of $i = 2$ can be proved analogously.

□

C.1. The kernel concentrations.

Before proving the equivalence between sharpness and parameter norms, we first need to bound the difference between the poluation kernels and empirical kernels.

Lemma C.3. *Let $\mathcal{H} = \{h(\cdot;\theta) : \theta \in \Theta\}$. Denote by $\omega : [0,\infty) \mapsto [0,\infty)$ a modulus of continuity of h in the sense $\sup_{x\in\mathcal{X}} |h(x;\theta_1) - h(x;\theta_2)| \leqslant \omega(\|\theta_1 - \theta_2\|)$. Suppose that $\forall\, \theta \in \Theta$, $h(X;\theta)$ is mean zero and sub-exponential with $\|h(X;\theta)\|_{\psi_1} \leqslant K$. Let N_ε be the covering number of Θ with respect to $\|\cdot\|$. Then for any $\delta \in (0,1)$, w.p. at least $1-\delta$ it holds that*

$$\sup_{\theta\in\Theta} \left| \frac{1}{n}\sum_{i=1}^n h(x_i;\theta) \right| \lesssim \omega(\varepsilon) + K \max\left(\frac{\log(N_\varepsilon/\delta)}{n}, \sqrt{\frac{\log(N_\varepsilon/\delta)}{n}} \right).$$

Proof. Let Θ_ε be an ε-cover of Θ. For any $\theta \in \Theta$, let θ' be an element in Θ_ε such that $\|\theta - \theta'\| \leqslant \varepsilon$. Then,

$$\begin{aligned}
\sup_{\theta\in\Theta} \left| \frac{1}{n}\sum_{i=1}^n h(x_i;\theta) \right| &= \sup_{\theta\in\Theta} \left| \frac{1}{n}\sum_{i=1}^n \left(h(x_i;\theta) - h(x_i;\theta') + h(x_i;\theta') \right) \right| \\
&\leqslant \omega(\varepsilon) + \sup_{\theta\in\Theta_\varepsilon} \left| \frac{1}{n}\sum_{i=1}^n h(x_i;\theta) \right|. \qquad (28)
\end{aligned}$$

Since $h(X;\theta)$ is sub-exponential with $\|h(X;\theta)\|_{\psi_1} \leqslant K$. By the Bernstein inequality (Theorem A.10), it holds for any $\theta \in \Theta_\varepsilon$ that

$$\mathbb{P}\left\{ \left| \frac{1}{n}\sum_{i=1}^n h(x_i;\theta) \right| \geqslant t \right\} \leqslant 2e^{-Cn\min(\frac{t}{K}, \frac{t^2}{K^2})}.$$

Taking the union bound over Θ_ε leads to

$$\mathbb{P}\left\{ \sup_{\theta\in\Theta_\varepsilon} \left| \frac{1}{n}\sum_{i=1}^n h(x_i;\theta) \right| > t \right\} \leqslant |\Theta_\varepsilon| 2e^{-Cn\min(\frac{t}{K}, \frac{t^2}{K^2})}.$$

Since $|\Theta_\epsilon| = N_\varepsilon$, the above implies w.p. $1-\delta$ that

$$\sup_{\theta\in\Theta_\varepsilon} |\frac{1}{n}\sum_{i=1}^n h(x_i;\theta)| \lesssim K \max\left(\frac{\log(N_\varepsilon/\delta)}{n}, \sqrt{\frac{\log(N_\varepsilon/\delta)}{n}} \right).$$

Substituting it into (28) completes the proof. □

Lemma C.4. *Let Z be a mean-zero and sub-Gaussian random variable. Assume that q is L_q-Lipschitz and $q(c) = 0$ for some c. Then, $q(Z)$ is also sub-Gaussian with $\|q(Z)\|_{\psi_2} \lesssim L_q(\|Z\|_{\psi_2} + |c|)$.*

Proof. By the property of sub-Gaussian random variable, we have $\mathbb{E}[e^{Z^2/\|Z\|_{\psi_2}^2}] \leqslant 2$. Then,

$$\begin{aligned}
\mathbb{E}[e^{q(Z)^2/(L_q^2\|Z-c\|_{\Psi_2}^2)}] &= \mathbb{E}[e^{(q(Z)-q(c))^2/(L_q^2\|Z-c\|_{\Psi_2}^2)}] \\
&\leqslant \mathbb{E}[e^{L_q^2|Z-c|^2/(L_q^2\|Z-c\|_{\Psi_2}^2)}] = \mathbb{E}[e^{|Z-c|^2/\|Z-c\|_{\Psi_2}^2}] \leqslant 2.
\end{aligned}$$

Hence, we have $\|q(Z)\|_{\psi_2} \leqslant L_q\|Z-c\|_{\psi_2} \lesssim L_q(\|Z\|_{\psi_2} + |c|)$, where the last inequality is due to that $\|\cdot\|_{\psi_2}$ is a norm. □

Lemma C.5. *Let $p : \mathbb{R} \mapsto \mathbb{R}$ be L_p-Lipschitz and $p(0) = 0$ and $q : \mathbb{R} \mapsto \mathbb{R}$ be L_q-Lipschitz and $q(0) = 0$. Let $Z_{u,v} = p(u^TX)q(v^TX) - \mathbb{E}[p(u^TX)q(v^TX)]$. Then, for any $u, v \in \mathbb{S}^{d-1}$, $Z_{u,v}(X)$ is a mean-zero and satisfies*

$$\begin{aligned}
\|Z_{u,v}(X)\|_{\psi_1} &\lesssim L_pL_q\|X\|_{\psi_2}^2 \\
|Z_{u,v}(X) - Z_{u',v'}(X)| &\lesssim L_pL_q\|X\|^2(\|u-u'\| + \|v-v'\|).
\end{aligned}$$

Proof. We first have

$$\begin{aligned}\|Z_{u,v}(X)\|_{\psi_1} &\leqslant C\|p(u^TX)q(v^TX)\|_{\psi_1} \leqslant C\|p(u^TX)\|_{\psi_2}\|q(v^TX)\|_{\psi_2}\\ &\leqslant CL_pL_q\|u^TX\|_{\psi_2}\|v^TX\|_{\psi_2} \leqslant CL_pL_q\|X\|_{\psi_2}^2,\end{aligned}$$

where the first step follows from the centering inequality (16); the second step is due to Lemma A.9; the third inequality follows from Lemma C.4.

Let $J_{u,v} = p(u^TX)q(v^TX)$. Then,

$$\begin{aligned}|J_{u,v} - J_{u',v'}| &= |J_{u,v} - J_{u,v'}| + |J_{u,v'} - J_{u',v'}|\\ &= |p(u^TX)||q(v^TX) - q(v'\cdot X)| + |q(v'\cdot X)||p(u^TX) - p(u'\cdot X)|\\ &\leqslant CL_pL_q\|X\|^2(\|v - v'\| + \|u - u'\|).\end{aligned}$$

Analogously, we can prove that $Z_{u,v}(X)$ satisfies the same Lipschitz condition. □

Now we are ready to bound the difference between population kernels and the corresponding empirical kernels.

Lemma C.6. *Suppose $n \gtrsim d$. For any $\delta \in (0,1)$, w.p. at least $1-\delta$, it holds for any $i = 1, 2$ that*

$$\sup_{u,v\in\mathbb{S}^{d-1}} |\varphi_i(u^Tv) - \hat{\varphi}_i(u,v)| \lesssim \min\left\{\sqrt{\frac{d\log(n/\delta)}{n}}, \frac{d\log(n/\delta)}{n}\right\} =: r_n.$$

Proof. Recall $\Omega = \mathbb{S}^{d-1}\otimes\mathbb{S}^{d-1}$ and let $\|(u,v) - (u',v')\|_\Omega = \|u-u'\| + \|v-v'\|$ for any $(u,v),(u',v') \in \Omega$.

The case of φ_1. Let $h(x;\theta) = \sigma(u^Tx)\sigma(v^Tx)$. By Lemma C.5 and noticing that σ is Lipschitz continuous, we have

$$\begin{aligned}\|h(X;\theta)\|_{\psi_1} &\lesssim 1\\ |h(X;\theta_1) - h(X;\theta_2)| &\lesssim \|X\|^2\|\theta_1 - \theta_2\|_\Omega \lesssim d\|\theta_1 - \theta_2\|_\Omega.\end{aligned} \tag{29}$$

By Lemma C.1, $N_\epsilon \leqslant (6/\epsilon)^{2d}$. Then applying Lemma C.3 gives w.p. at least $1-\delta$ it holds that

$$\begin{aligned}\sup_{u,v\in\mathbb{S}^{d-1}} |\hat{\varphi}_1(u,v) - \varphi_1(u,v)| &\lesssim d\epsilon + \max\left\{\frac{\log(N_\epsilon/\delta)}{n}, \sqrt{\frac{\log(N_\epsilon/\delta)}{n}}\right\}\\ &\lesssim \frac{d}{n} + \max\left\{\frac{d\log(n/\delta)}{n}, \sqrt{\frac{d\log(n/\delta)}{n}}\right\} \leqslant 2r_n,\end{aligned}$$

where we take $\epsilon = 1/n$.

The case of φ_2. For φ_2, the major challenge comes the discontinuity of $\sigma'(\cdot)$. Fortunately, the concentration is still possible since σ' is discontinuous only at the origin. Note that $\sigma'(\cdot)$ is exactly the Heaviside step function and hence, we will write $H(t) = \sigma'(t)$ for convenience.

- **Step 1: smoothing the kernels.** Define two smoothed Heaviside step functions:

$$H_\beta^-(t) = \begin{cases}1 & \text{if } t \geqslant \beta\\ \frac{1}{\beta}t & \text{if } 0 \leqslant t \leqslant \beta\\ 0 & \text{if } t < 0.\end{cases} \qquad H_\beta^+(t) = \begin{cases}1 & \text{if } t \geqslant 0\\ \frac{1}{\beta}t + 1 & \text{if } -\beta \leqslant t \leqslant 0\\ 0 & \text{if } t < -\beta,\end{cases}$$

  where $\beta \in \mathbb{R}_+$ control the degree of smoothing. Then, we have that H_β^+, H_β^{-1} are both $\frac{1}{\beta}$-Lipschitz and $0 \leqslant H_\beta^-(t) \leqslant H(t) \leqslant H_\beta^+(t) \leqslant 1$. An illustration of these three functions are provided in Figure 4.

  Correspondingly, define

$$\varphi_{2,\beta}^+(u^Tv) = \mathbb{E}_x[H_\beta^+(u^Tx)H_\beta^+(v^Tx)], \quad \varphi_{2,\beta}^-(u^Tv) = \mathbb{E}_x[H_\beta^-(u^Tx)H_\beta^-(v^Tx)].$$

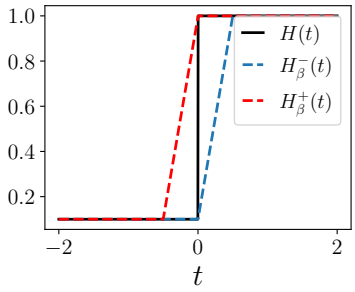

Figure 4. An visual comparison among H, H_β^+, and H_β^-.

We first have

$$\begin{aligned}
|\varphi_{2,\beta}^+(u^Tv) - \varphi_2(u^Tv)| &= |\,\mathbb{E}_x[H_\beta^+(u^Tx)H_\beta^+(v^Tx)] - \mathbb{E}_x[H(u^Tx)H(v^Tx)]| \\
&\leqslant |\,\mathbb{E}_x[H_\beta^+(u^Tx)(H_\beta^+(v^Tx) - H(v^Tx))] + \mathbb{E}_x[(H_\beta^+(u^Tx) - H(u^Tx))H(v^Tx)]| \\
&\overset{(i)}{\leqslant} \mathbb{E}_x\,|H_\beta^+(u^Tx) - H(u^Tx)| + \mathbb{E}_x\,|H_\beta^+(v^Tx) - H(v^Tx)| \\
&\overset{(ii)}{=} 2\int_0^\beta (1 - \frac{t}{\beta})p_d(t)\,\mathrm{d}t \overset{(iii)}{\lesssim} \beta,
\end{aligned}$$

where (i) follows from the boundedness of H, H_β^+, (ii) follows from the the fact that the input distribution is $\rho = \mathrm{Unif}(\sqrt{d}\mathbb{S}^{d-1})$ and $p_d(\cdot)$ denotes the distribution of X_1 for $X \sim \mathrm{Unif}(\sqrt{d}\mathbb{S}^{d-1})$.

Similarly, we can obtain

$$\sup_{u,v\in\mathbb{S}^{d-1}} |\varphi_{2,\beta}^-(u^Tv) - \varphi_2(u^Tv)| \lesssim \beta. \tag{30}$$

- **Step 2: concentration through smoothing.** Let $h_\beta^-(x;\theta) = H_\beta^-(x^Tu)H_\beta^-(x^Tv)$. Note that for any $x \in \sqrt{d}\mathbb{S}^{d-1}$, we have

$$\begin{aligned}
|h_\beta^-(x;\theta_1) - h_\beta^-(x;\theta')| &\leqslant |H_\beta^-(x^Tu)H_\beta^-(x^Tv) - H_\beta^-(x^Tu)H_\beta^-(x^Tv')| + |H_\beta^-(x^Tu)H_\beta^-(x^Tv') - H_\beta^-(x^Tu')H_\beta^-(x^Tv')| \\
&\leqslant |H_\beta^-(x^Tv) - H_\beta^-(x^Tv')| + |H_\beta^-(x^Tu) - H_\beta^-(x^Tu')| \\
&\leqslant \frac{1}{\beta}(|x^Tv - x^Tv'| + |x^Tu - x^Tu'|) \leqslant \frac{\sqrt{d}}{\beta}(\|v - v'\| + \|u - u'\|).
\end{aligned}$$

Hence, $h(x;\cdot)$ is $\frac{\sqrt{d}}{\beta}$-Lipschitz in $(\Omega, \|\cdot\|_\Omega)$. In addition, since $\sup_t |H_\beta^-(t)| \leqslant 1$, we have $\|h(X;\theta)\|_{\psi_1} \lesssim 1$ for any $\theta \in \Omega$. Then, applying Lemma C.3 gives that w.p. at least $1-\delta$ it holds that

$$\sup_{u,v\in\mathbb{S}^{d-1}} |\frac{1}{n}\sum_{i=1}^n H_\beta^-(u^Tx)H_\beta^-(v^Tx) - \varphi_{2,\beta}^-(u^Tv)| \lesssim \frac{\sqrt{d}\epsilon}{\beta} + \max\left\{\sqrt{\frac{d\log(1/(\epsilon\delta))}{n}}, \frac{d\log(1/(\epsilon\delta))}{n}\right\}$$

By $H(t) \geqslant H_\beta^-(t)$ and the above inequality, we have

$$\begin{aligned}
\hat{\varphi}_2(u^Tv) = \frac{1}{n}\sum_{i=1}^n H(u^Tx_i)H(v^Tx_i) &\geqslant \frac{1}{n}\sum_{i=1}^n H_\beta^-(u^Tx_i)H_\beta^-(v^Tx_i) \\
&\geqslant \varphi_{2,\beta}^-(u^Tv) - C\left(\frac{\sqrt{d}\varepsilon}{\beta} + \max\left\{\sqrt{\frac{d\log(1/(\epsilon\delta))}{n}}, \frac{d\log(1/(\epsilon\delta))}{n}\right\}\right) \\
&\geqslant \varphi_2(u^Tv) - C\left(\beta + \frac{\sqrt{d}\varepsilon}{\beta} + \max\left\{\sqrt{\frac{d\log(1/(\epsilon\delta))}{n}}, \frac{d\log(1/(\epsilon\delta))}{n}\right\}\right),
\end{aligned}$$

where the last step uses (30). Optimizing β gives

$$\hat{\varphi}_2(u^T v) \geqslant \varphi_2(u^T v) - C\left((\sqrt{d}\varepsilon)^{1/2} + \max\left\{\sqrt{\frac{d\log(1/(\epsilon\delta))}{n}}, \frac{d\log(1/(\epsilon\delta))}{n}\right\}\right).$$

Taking $\varepsilon = 1/n$ and applying $n \geqslant d$, the above inequality can be simplified as

$$\hat{\varphi}_2(u^T v) \geqslant \varphi_2(u^T v) - C\max\left\{\sqrt{\frac{d\log(n/\delta)}{n}}, \frac{d\log(n/\delta)}{n}\right\}. \tag{31}$$

Similarly, by utilizing H_β^+ and $\varphi_{2,\beta}^+$, we can prove

$$\hat{\varphi}_2(u^T v) \leqslant \varphi_2(u^T v) + C\max\left\{\sqrt{\frac{d\log(n/\delta)}{n}}, \frac{d\log(n/\delta)}{n}\right\}. \tag{32}$$

Combining (31) and (32), we complete the proof.

□

C.2. Proof of Theorem 4.1

The expression of Fisher matrix. Notice that $\nabla_{w_j} f(x;\theta) = a_j\sigma'(w_j^T x)x, \nabla_{a_j} f(x;\theta) = \sigma(w_j^T x)$. Then the Fisher matrix is given by

$$G(\theta) = \begin{pmatrix} F_{1,1} & F_{1,2} & \dots & F_{1,m} \\ F_{2,1} & F_{2,2} & \dots & F_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ F_{m,1} & F_{m,2} & \dots & F_{m,m} \end{pmatrix} \in \mathbb{R}^{m(d+1)\times m(d+1)}, \tag{33}$$

where for any $j, k \in [m]$ the submatrix $F_{j,k} \in \mathbb{R}^{(d+1)\times(d+1)}$ is given by

$$F_{j,k} = \begin{pmatrix} \hat{\mathbb{E}}[\sigma(w_j^T x)\sigma(w_k^T x)] & \hat{\mathbb{E}}[\sigma(w_j^T x)a_k\sigma'(w_k^T x)x^T] \\ \hat{\mathbb{E}}[\sigma(w_k^T x)a_j\sigma'(w_j^T x)x] & a_j a_k\hat{\mathbb{E}}[\sigma'(w_j^T x)\sigma'(w_k^T x)xx^T] \end{pmatrix}. \tag{34}$$

Proof of Theorem 4.1. We consider the case of trace, Frobenius norm, and the spectral norm separately. Specifically, combining Proposition C.7, C.15, and C.11, we complete the proof. The proofs of these propositions are provided in the subsequent sections.

C.2.1. The trace of Fisher matrix

Proposition C.7 (The trace)**.** *Recall that* $N(d,\delta) := \inf\{n : d\log(1/\delta)/n \leqslant 1\}$. *For any* $\delta \in (0,1)$, *let* $n \geqslant N(d,\delta)$. *Then, w.p.* $1-\delta$, *we have,* $\mathrm{Tr}(G(\theta)) \sim \sum_j(\|w_j\|^2 + da_j^2)$.

Proof. It is easy to show that

$$\mathrm{Tr}(G(\theta)) = \sum_{j=1}^m \left(\|w_j\|^2\hat{\varphi}_1(\hat{w}_j,\hat{w}_j) + da_j^2\hat{\varphi}_2(\hat{w}_j,\hat{w}_j)\right). \tag{35}$$

By Lemma C.6, we have $\hat{\varphi}_i(u,u) \sim \varphi_i(u,u) - r_n \sim 1 - o(1)$, where the last inequality is due to Lemma C.2 and the condition $n \geqslant N(d,\delta)$. Plugging this into (35), we complete the proof. □

C.2.2. THE FROBENIUS NORM OF FISHER MATRIX

To help the estimate of Frobenius norm, we define for $u, v \in \mathbb{S}^{d-1}$ that

$$b_{u,v} = \hat{\mathbb{E}}[\sigma(u^T x)\sigma'(v^T x)x] \in \mathbb{R}^d, \quad A_{u,v} = \hat{\mathbb{E}}[\sigma'(u^T x)\sigma'(v^T x)xx^T] \in \mathbb{R}^{d\times d}.$$

Then by (33), we have

$$\|G(\theta)\|_F^2 = \sum_{j,k=1}^m \left(\|w_j\|^2\|w_k\|^2\hat{\varphi}_1(\hat{w}_j,\hat{w}_k)^2 + a_j^2a_k^2\|A_{\hat{w}_j,\hat{w}_k}\|_F^2 + a_j^2\|w_k\|^2\|b_{\hat{w}_k,\hat{w}_j}\|_2^2 + a_k^2\|w_j\|^2\|b_{\hat{w}_j,\hat{w}_k}\|_2^2\right). \quad (36)$$

Next, we bound each term of the right hand side separately.

Lemma C.8. *For any $\delta \in (0,1)$, if $n \gtrsim N(d,\delta)$, w.p. $1-\delta$ it holds that $\sup_{u,v\in\mathbb{S}^{d-1}} \|b_{u,v}\|_2 \lesssim 1$.*

Proof. Note that for any $u, v \in \mathbb{S}^{d-1}$,

$$\begin{aligned}
\|b_{u,v}\| &= \sup_{\|w\|=1} w^T b_{u,v} = \sup_{\|w\|=1} \hat{\mathbb{E}}[\sigma(u^T x)\sigma'(v^T x)w^T x] \\
&\lesssim \sup_{\|w\|=1} \hat{\mathbb{E}}[\sigma(u^T x)w^T x] \leqslant \sup_{\|w\|=1} \sqrt{\hat{\mathbb{E}}[\sigma^2(u^T x)]}\sqrt{\hat{\mathbb{E}}[|w^T x|^2]} \qquad \text{(Cauchy-Schwartz)} \\
&= \sup_{w\in\mathbb{S}^{d-1}} \sqrt{\hat{\varphi}_1(u,u)}\sqrt{w^T\hat{\Sigma}_n w} = \sqrt{\hat{\varphi}_1(u,u)}\lambda_{\max}(\hat{\Sigma}_n) \lesssim 1,
\end{aligned}$$

where the last steps follows from Lemma C.6 and Lemma A.14. □

Lemma C.9. *For any $\delta \in (0,1)$, if $n \gtrsim d + \log(1/\delta)$, then w.p. $1-\delta$ it holds for any $u, v \in \mathbb{S}^{d-1}$ that*

$$\sqrt{d}\hat{\varphi}_2(u,v) \leqslant \|A_{u,v}\|_F \lesssim \sqrt{d}.$$

Proof. **Upper bound.** We first prove a more general result. Let $a \in \mathbb{R}^n$ with $\sup_{i\in[n]}|a_i| \lesssim 1$ and $Q_a = \frac{1}{n}\sum_{i=1}^n a_i x_i x_i^T$. Then,

$$\|\frac{1}{n}\sum_{i=1}^n a_i x_i x_i^T\|_F^2 = \frac{1}{n^2}\sum_{i,j=1}^n a_i a_j (x_i^T x_j)^2 \lesssim \frac{1}{n^2}\sum_{i,j=1}^n (x_i^T x_j)^2 = \|\frac{1}{n}\sum_{i=1}^n x_i x_i^T\|_F^2 = \|\hat{\Sigma}_n\|_F^2.$$

By Lemma A.14, w.p. $1-\delta$ that $\|\hat{\Sigma}_n\|_F \lesssim \sqrt{d}$. Thus, $\|Q_a\|_F \leqslant \sqrt{d}$ for any $\|a\|_\infty \lesssim 1$. Notice that we can rewrite $A_{u,v}$ as $A_{u,v} = \frac{1}{n}\sum_{i=1}^n \sigma'(u^T x_i)\sigma'(v^T x_i)x_i x_i^T$, with $|\sigma'(u^T x_i)\sigma'(v^T x_i)| \lesssim 1$. Thus, $\|A_{u,v}\|_F \leqslant \sqrt{d}$.

Lower bound. Now we consider the lower bound:

$$\begin{aligned}
\|A_{u,v}\|_F &\geqslant \frac{1}{\sqrt{d}}\operatorname{Tr}(A_{u,v}) = \frac{1}{\sqrt{d}}\sum_{j=1}^d \hat{\mathbb{E}}[\sigma'(u^T x)\sigma'(v^T x)x_j^2] \\
&= \sqrt{d}\hat{\mathbb{E}}[\sigma'(u^T x)\sigma'(v^T x)] = \sqrt{d}\,\hat{\varphi}_2(u,v).
\end{aligned}$$

□

Lemma C.10. *For any $\delta \in (0,1)$, if $n \gtrsim N(d,\delta)$, then w.p. at least $1-\delta$ it holds for $i = 1, 2$ that*

$$\sum_{j,k=1}^m \alpha_j^2\alpha_k^2\hat{\varphi}_i(\hat{w}_j,\hat{w}_k) \sim \sum_j \alpha_j^2.$$

Proof. WLOG, assume $\sum_j \alpha_j^2 = 1$ and let $\Delta_{j,k} = \varphi_1(\hat{w}_j,\hat{w}_k) - \hat{\varphi}_1(\hat{w}_j,\hat{w}_k)$. By Lemma C.6, w.p. at least $1-\delta$ it holds that $\sup_{j,k\in[m]}|\Delta_{j,k}| \leqslant r_n$, where r_n is defined in Lemma C.6. Hence,

$$\sum_{j,k=1}^m \alpha_j^2\alpha_k^2\hat{\varphi}_1(\hat{w}_j,\hat{w}_k)^2 = \sum_{j,k=1}^m \alpha_j^2\alpha_k^2(\varphi_1(\hat{w}_j,\hat{w}_k) + \Delta_{j,k})^2$$

$$= \sum_{j,k=1}^{m} \alpha_j^2\alpha_k^2\varphi_1(\hat{w}_j,\hat{w}_k)^2 + 2\sum_{j,k=1}^{m} \alpha_j^2\alpha_k^2\varphi_1(\hat{w}_j,\hat{w}_k)\Delta_{j,k} + \sum_{j,k=1}^{m} \alpha_j^2\alpha_k^2\Delta_{j,k}^2$$
$$\gtrsim (1+O(r_n^2))\sum_{j,k=1}^{m} \alpha_j^2\alpha_k^2 + O(r_n)\sum_{j,k=1}^{m} \alpha_j^2\alpha_k^2, \tag{37}$$

where the last step follows the third conclusion in Lemma C.2 and the fact that $\sup_t |\varphi_1(t)| \lesssim 1$.

Taking n to be large enough, we complete the proof. The case of $i = 2$ follows the same proof procedure. □

Now we are ready to prove the main proposition.

Proposition C.11 (The Frobenius norm)**.** *For any $\delta \in (0,1)$, let $n \geqslant N(d,\delta)$. Then, w.p. $1-\delta$, we have $\|G(\theta)\|_F \sim \sum_j(\|w_j\|^2 + \sqrt{d}a_j^2)$.*

Proof. **Lower bound.** By (36), w.p. at least $1-\delta$ that

$$\begin{aligned}
\|G(\theta)\|_F^2 &\geqslant \sum_{j,k=1}^{m}\left(\|w_j\|^2\|w_k\|^2\hat{\varphi}_1(\hat{w}_j,\hat{w}_k)^2 + a_j^2a_k^2A_{\hat{w}_j,\hat{w}_k}^2\right)\\
&\geqslant \sum_{j,k=1}^{m}\|w_j\|^2\|w_k\|^2\hat{\varphi}_1(\hat{w}_j,\hat{w}_k)^2 + d\sum_{j,k=1}^{m}a_j^2a_k^2\hat{\varphi}_2(\hat{w}_j,\hat{w}_k)^2 \qquad \text{(Use Lemma C.9)}\\
&\gtrsim (\sum_{j}^{m}\|w_j\|^2)^2 + (\sqrt{d}\sum_j a_j^2)^2 \qquad \text{(Use Lemma C.10)}\\
&\geqslant \frac{1}{2}\left(\sum_j\|w_j\|^2 + \sqrt{d}\sum_j a_j^2\right)^2 \qquad (\,(x^2+y^2) \geqslant (x+y)^2/2).
\end{aligned} \tag{38}$$

Upper bound. By Lemma C.8 and C.9, we have w.p. $1-\delta$ that $\|b_{\hat{w}_j,\hat{w}_k}\| \lesssim 1$, $\|A_{\hat{w}_j,\hat{w}_k}\|_F \lesssim \sqrt{d}$. Substituting it into (36) gives

$$\begin{aligned}
\|G(\theta)\|_F &\leqslant \sum_{j,k=1}^{m}(\|w_j\|^2\|w_k\|^2\hat{\varphi}_1(\hat{w}_j,\hat{w}_k)^2 + \sqrt{d}a_j^2\|w_k\|^2 + \sqrt{d}a_k^2\|w_j\|^2 + da_j^2a_k^2)\\
&\lesssim \sum_{j,k=1}^{m}(\|w_j\|^2\|w_k\|^2 + \sqrt{d}a_j^2 + \sqrt{d}a_k^2 + da_j^2a_k^2) \qquad \text{(Use Lemma C.10)}\\
&= \sum_{j,k}(\|w_j\|^2 + \sqrt{d}a_j^2)(\|w_k\|^2 + \sqrt{d}a_k^2) = (\sum_j(\|w_j\|^2 + \sqrt{d}a_j^2))^2.
\end{aligned}$$

□

C.2.3. THE SPECTRAL NORM

To control the spectral norm, we need again to handle the discontinuity of σ' at the origin. Define

$$\begin{gathered}
\phi_\beta^+(u^Tv) = \mathbb{E}_x[H_\beta^+(u^Tx)\sigma(v^Tx)], \quad \phi_\beta^-(u^Tv) = \mathbb{E}_x[H_\beta^-(u^Tx)\sigma(v^Tx)]\\
\phi(u^Tv) = \mathbb{E}[H(u^Tx)\sigma(v^Tx)].
\end{gathered} \tag{39}$$

Lemma C.12. *$|\phi_\beta^+(u^Tv) - \phi(u^Tv)| \lesssim \beta$ and $|\phi_\beta^-(u^Tv) - \phi(u^Tv)| \lesssim \beta$.*

Proof. Note that

$$|\phi_\beta^-(u^Tv) - \phi(u^Tv)| = |\,\mathbb{E}[H_\beta^-(u^Tx)\sigma(v^Tx)] - \mathbb{E}[H(u^Tx)\sigma(v^Tx)]|$$

$$= \mathbb{E}[(H_\beta^-(u^Tx) - H(u^Tx))\sigma(v^Tx)] \leqslant \sqrt{\mathbb{E}[\sigma^2(v^Tx)]}\sqrt{\mathbb{E}[(H_\beta^-(u^Tx) - H(u^Tx))^2]}$$
$$\lesssim \sqrt{\mathbb{E}_z[|H_\beta^-(z) - H(z)|^2]} = \sqrt{\int_0^\beta (1-\frac{z}{\beta})^2 p_d(z)\,\mathrm{d}z} \qquad \text{(Let } z = u^Tx\text{)}$$
$$\lesssim \sqrt{\int_0^\beta (1-\frac{z}{\beta})^2\,\mathrm{d}z} \lesssim \beta, \tag{40}$$

where p_d is the density function of u^TX for $X \sim \mathrm{Unif}(\sqrt{d}\mathbb{S}^{d-1})$ and we use the fact that $\sup_z |p_d(z)| \lesssim 1$. Similarly, we can prove the case of ϕ_β^+. □

Lemma C.13. *For any $\delta \in (0,1)$, if $n \gtrsim N(d,\delta)$, then w.p. at least $1-\delta$ that*

$$\sup_{u,v\in\mathbb{S}^{d-1}} |\hat{\mathbb{E}}[\sigma'(u^Tx)v^Tx] - \mathbb{E}[\sigma'(u^Tx)v^Tx]| \lesssim r_n.$$

Proof. The proof is essentially similar to the proof of Lemma C.6.

- **Step 1.** Note that $\|H_\beta^+(X)\|_{\psi_2} \lesssim 1$ and $\|\sigma(v^TX)\|_{\psi_2} \lesssim 1$. By Lemma A.9, we have

$$\|H_\beta^+(X)\sigma(v^TX)\|_{\psi_1} \leqslant \|H_\beta^+(X)\|_{\psi_2}\|\sigma(v^TX)\|_{\psi_2} \lesssim 1.$$

  The fact that H_β is $\frac{1}{\beta}$-Lipschitz continuous and σ is 1-Lipschitz implies that for any $x \in \sqrt{d}\mathbb{S}^{d-1}$, $J_x(u,v) := H_\beta^+(u^Tx)\sigma(v^Tx)$ is $\frac{d}{\beta}$ Lipschitz with respect to the metric $\|(u,v)-(u',v')\|_\Omega := \|u-u'\| + \|v-v'\|$. In addition, by Lemma C.1, the covering number of $\Omega = \mathbb{S}^{d-1}\otimes\mathbb{S}^{d-1}$ with respect to this metric is $N_\epsilon = (6/\epsilon)^{2d}$. Then, by Lemma C.3, we have

$$\sup_{u,v\in\mathbb{S}^{d-1}} |\hat{\mathbb{E}}_x[H_\beta^+(u^Tx)\sigma(v^Tx)] - \phi_\beta^+(u^Tv)| \lesssim \frac{d}{\beta}\epsilon + \max\{\sqrt{\frac{d\log(1/(\epsilon\delta))}{n}}, \frac{d\log(1/(\epsilon\delta))}{n}\}. \tag{41}$$

  Similarly, we can obtain the following holds w.p. at least $1-\delta$,

$$\sup_{u,v\in\mathbb{S}^{d-1}} |\hat{\mathbb{E}}_x[H_\beta^-(u^Tx)\sigma(v^Tx)] - \phi_\beta^-(u^Tv)| \lesssim \frac{d}{\beta}\epsilon + \max\{\sqrt{\frac{d\log(1/(\epsilon\delta))}{n}}, \frac{d\log(1/(\epsilon\delta))}{n}\}. \tag{42}$$

- **Step 2.** Noting $t = \sigma(t) - \sigma(-t)$ for any $t \in \mathbb{R}$, we have

$$\begin{aligned}\hat{\mathbb{E}}[\sigma'(u^Tx)v^Tx] &= \hat{\mathbb{E}}[H(u^Tx)\sigma(v^Tx)] - \hat{\mathbb{E}}[H(u^Tx)\sigma(-v^Tx)] \\ &\geqslant \hat{\mathbb{E}}[H_\beta^-(u^Tx)\sigma(v^Tx)] - \hat{\mathbb{E}}[H_\beta^+(u^Tx)\sigma(-v^Tx)] \\ &\geqslant \phi_\beta^-(u^Tv) - \phi_\beta^+(-u^Tv) - C\left(\frac{d}{\beta}\epsilon + \max\{\sqrt{\frac{d\log(1/(\epsilon\delta))}{n}}, \frac{d\log(1/(\epsilon\delta))}{n}\}\right),\end{aligned} \tag{43}$$

  where the last inequality follows from (41) and (42).

- **Step 3.** Applying Lemma C.12 to (43) gives

$$\begin{aligned}\hat{\mathbb{E}}[\sigma'(u^Tx)v^Tx] &\geqslant \phi(u^Tv) - \phi(-u^Tv) - C\left(\beta + \frac{d}{\beta}\epsilon + \max\{\sqrt{\frac{d\log(1/(\epsilon\delta))}{n}}, \frac{d\log(1/(\epsilon\delta))}{n}\}\right) \\ &= \mathbb{E}[\sigma'(u^Tx)v^Tx] - C\left(\beta + \frac{d}{\beta}\epsilon + \max\{\sqrt{\frac{d\log(1/(\epsilon\delta))}{n}}, \frac{d\log(1/(\epsilon\delta))}{n}\}\right).\end{aligned}$$

Optimizing β and taking $\epsilon = 1/n$, we obtain

$$\begin{aligned}\hat{\mathbb{E}}[\sigma'(u^T x)v^T x] &\geqslant \mathbb{E}[\sigma'(u^T x)v^T x] - C\left(d^{1/2}\epsilon^{1/2} + \max\{\sqrt{\frac{d\log(1/(\epsilon\delta))}{n}}, \frac{d\log(1/(\epsilon\delta))}{n}\}\right)\\ &\geqslant \mathbb{E}[\sigma'(u^T x)v^T x] - Cr_n.\end{aligned}$$

Analogously, we can prove the other side inequality.

□

Let $x \in \mathbb{R}^d$, we use $x_{,k}$ to denote the k-th coordinate of x, which is distinguished from x_k, denoting the k-th sample in the training set. In addition, we use $\{e_k\}_{k=1}^d$ to denote the standard basis of $\mathbb{R}^d$.

Lemma C.14. $\sum_{k=1}^d (\mathbb{E}[\sigma'(u^T x)x_{,k}])^2 \gtrsim 1$

Proof. Note that $x_{,k} = g(e_k^T x)$ with $g(x) = x$. Let $g(x) = \sum_s G_s h_s(x)$ be the Hermite expansion of g. By (13), we have $G_k = 1$ for $k = 1$ and 0 otherwise. Then Lemma A.1 gives that for any $u \in \mathbb{S}^{d-1}$,

$$\mathbb{E}[\sigma'(u^T x)x_{,k}] = \mathbb{E}[\sigma'(u^T x)g(e_k^T x)] = \sum_s \beta_s G_s (u^T e_k)^s = \beta_1 u_k.$$

Hence, $\sum_{k=1}^d (\mathbb{E}[\sigma'(u^T x)x_{,k}])^2 = \sum_{k=1}^d (\beta_1 u_k)^2 = \beta_1^2 \|u\|^2 = \beta_1^2 \gtrsim 1$. □

Proposition C.15 (The spectral norm)**.** *For any $\delta \in (0,1)$, if $n \gtrsim dN(d,\delta)$, w.p. $1-\delta$ we have $\|G_\theta\|_2 \sim \sum_{j=1}^m (w_j^2 + a_j^2)$.*

Proof. Let $\Phi = (\nabla f(x_1;\theta), \nabla f(x_2;\theta), \ldots, \nabla f(x_n;\theta)) \in \mathbb{R}^{p\times n}$. Then, $G(\theta) = \Phi\Phi^T \in \mathbb{R}^{p\times p}$. Then

$$\begin{aligned}\|G(\theta)\|_2 &= \lambda_{\max}(\Phi\Phi^T) = \lambda_{\max}(\Phi^T\Phi) = \sup_{u\in\mathbb{S}^{n-1}} u^T\Phi^T\Phi u = \sup_{u\in\mathbb{S}^{n-1}} \|\Phi u\|^2\\ &= \sup_{\|u\|_{L^2(\hat{\rho})}=1} \sum_{j=1}^m \left((\hat{\mathbb{E}}[\sigma(w_j^T x)u(x)])^2 + a_j^2 \sum_{k=1}^d (\hat{\mathbb{E}}[\sigma'(w_j^T x)x_{,k}u(x)])^2\right), \end{aligned} \tag{44}$$

where $\hat{\rho} = \frac{1}{n}\sum_{i=1}^n \delta(x_i - \cdot)$.

Lower bound. Taking $u(x) \equiv 1$, we obtain

$$\begin{aligned}\|G(\theta)\|_2 &\geqslant \sum_{j=1}^m \left((\hat{\mathbb{E}}[\sigma(w_j^T x)])^2 + a_j^2 \sum_{k=1}^d (\hat{\mathbb{E}}[\sigma'(w_j^T x)x_{,k}])^2\right)\\ &= \sum_{j=1}^m \left(\|w_j\|^2 \hat{\varphi}_1(\hat{w}_j, \hat{w}_j) + a_j^2 \sum_{k=1}^d (\mathbb{E}[\sigma'(w_j^T x)x_{,k}])^2\right). \end{aligned} \tag{45}$$

Note that

- By Lemma C.6, w.p. at least $1-\delta$ that $\hat{\varphi}_1(u,u) \gtrsim 1$ for any $u \in \mathbb{S}^{d-1}$.
- In addition,

$$\begin{aligned}\sum_{k=1}^d (\hat{\mathbb{E}}[\sigma'(w_j^T x)e_k^T x])^2 &= \sum_{k=1}^d (\mathbb{E}[\sigma'(w_j^T x)e_k^T x] + O(r_n))^2 \qquad \text{(Lemma C.13)}\\ &\geqslant \sum_{k=1}^d (\mathbb{E}[\sigma'(w_j^T x)e_k^T x])^2 + O(r_n)\sum_{k=1}^d \mathbb{E}[\sigma'(w_j^T x)e_k^T x]\\ &\gtrsim 1 - O(\sqrt{d}r_n),\end{aligned}$$

where the last step uses Lemma C.14 and $|\sum_{k=1}^d \mathbb{E}[\sigma'(u^T x)e_k^T x]| = |\mathbb{E}[\sigma'(u^T x)(\sum_k e_k)x]| \leqslant \sqrt{o^T \mathbb{E}[xx^T]o} \leqslant \|o\| = \sqrt{d}$, where $o = (1,1,\ldots,1) \in \mathbb{R}^d$.

Plugging the above two estimates into (45), we obtain $\|G_\theta\| \gtrsim \sum_{j=1}^m (w_j^2 + a_j^2)$.

Upper bound. In addition, (44) also implies

$$\begin{aligned}\|G_\theta\| &\leqslant \sum_{j=1}^m \left(\sup_{\|u\|_{L^2(\hat{\rho})}=1} (\hat{\mathbb{E}}[\sigma(w_j^T x)e(x)])^2 + a_j^2 \sup_{\|u\|_{L^2(\hat{\rho})}=1} \sum_{k=1}^d (\hat{\mathbb{E}}[\sigma'(w_j^T x)x_{,k}u(x)])^2 \right) \\ &\leqslant \sum_{j=1}^m \left(\hat{\mathbb{E}}[\sigma(w_j^T x)^2] + a_j^2 \|\hat{\mathbb{E}}_x[\sigma'(w^T x)^2 x x^T]\|_2^2 \right) \end{aligned} \tag{46}$$

where the last step follows from the Cauchy-Schwarz inequality and Lemma A.13. Then, the proof of the upper bound is completed by plugging the following estimates into (46).

- By Lemma C.6, w.p. $1-\delta$ that $\hat{\mathbb{E}}[\sigma(w_j^T x)^2] = \|w_j\|^2 \hat{\varphi}_1(\hat{w}_j, \hat{w}_j) \lesssim \|w_j\|^2$.
- In addition,

$$\begin{aligned}\|\mathbb{E}_x[\sigma'(w_j^T x)^2 x x^T]\|_2 &= \sup_{v\in\mathbb{S}^{d-1}} v^T \mathbb{E}_x[\sigma'(w_j^T x)^2 x x^T] v \\ &= \mathbb{E}_x[\sigma'(w_j^T x)^2 (v^T x)^2] \lesssim \mathbb{E}_x[|v^T x|^2] \lesssim \sup_{v\in\mathbb{S}^{d-1}} v^T \mathbb{E}[x x^T] v = 1.\end{aligned}$$

Lastly, combining the lower and upper bound, we complete the proof. □

C.3. Proof of Theorem 4.3

Our proof needs the following path-norm based generalization bound:

Proposition C.16. *Suppose* $\sup_{x\in\mathcal{X}} |f^*(x)| \leqslant 1$ *and* $\gamma \geqslant 1$. *If* $\hat{\theta}$ *is a global minimum of* $\hat{\mathcal{R}}(\cdot)$ *satisfying* $\|\hat{\theta}\|_{\mathcal{P}} \leqslant \gamma$, *then* $\mathcal{R}(\hat{\theta}) \leqslant (\log^3(n) + \log(1/\delta))\frac{d\gamma^2}{n}$.

Proof. Let $\mathcal{F}_\gamma = \{f(\cdot;\theta) : \|\theta\|_{\mathcal{P}} \leqslant \gamma\}$. Then, it is easy to show that the worst-case Rademacher complexity (see Eq. (14)) $\mathfrak{R}_n(\mathcal{F}_\gamma) \lesssim \sqrt{d}\gamma/\sqrt{n}$. In addition,

$$|f(x)| = |\sum_j a_j \sigma(w_j^T x)| \leqslant \sum_j |a_j||w_j^T x| \leqslant \|\theta\|_{\mathcal{P}}\|x\| \leqslant \gamma\sqrt{d}.$$

Hence, the loss function $\phi(t) = t^2/2$ satisfying $|\phi''| \lesssim 1$ and $|\phi| \lesssim \gamma^2 d$. Then, by Theorem A.7, we have

$$\mathcal{R}(\hat{\theta}) \lesssim \frac{\log^3(n) d\gamma^2}{n} + \frac{d\gamma^2 \log(1/\delta)}{n}.$$

□

Proof of Theorem 4.3. For $\hat{\theta}_{\mathrm{sgd}}$, by Proposition 3.2 and Theorem 4.1 we have w.p. at least $1-\delta$ that

$$\frac{2}{\eta} \geqslant \mathrm{Tr}(G(\tilde{\theta}_{\mathrm{sgd}})) \sim \|\tilde{\theta}_{\mathrm{sgd}}\|_{1,d} \geqslant \sqrt{d}\|\tilde{\theta}_{\mathrm{sgd}}\|_{\mathcal{P}}. \tag{47}$$

Hence, $\|\hat{\theta}_{\mathrm{sgd}}\|_{\mathcal{P}} \lesssim 2/(\eta\sqrt{d})$. Applying Proposition C.16, we obtain

$$\mathcal{R}(\hat{\theta}_{\mathrm{sgd}}) \lesssim \frac{\log^3(n) + \log(1/\delta)}{n\eta^2}.$$

Similarly, we have $\|\hat{\theta}_{\mathrm{gd}}\|_{\mathcal{P}} \lesssim 2/\eta$. Applying Proposition C.16 completes the proof. □

D. Missing Proofs in Section 5

We first need the following lemma.

Lemma D.1. *Let X be a mean-zero with $\|X\|_{\psi_2} \lesssim 1$. Let $\Sigma = \mathbb{E}[XX^T]$ and $\hat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^n X_i X_i^T$ be the population and empirical covariance matrix, respectively. For any $\delta \in (0,1)$, if $n \gtrsim \log(d/\delta)$, we have w.p. $1-\delta$ the following holds for any $j,k \in [d]$.*

$$|(\hat{\Sigma}_n)_{j,k} - (\Sigma)_{j,k}| \lesssim \sqrt{\frac{\log(d/\delta)}{n}}.$$

Proof. First, for each $j,k \in [d]$, $(\hat{\Sigma}_n)_{j,k} = \frac{1}{n}\sum_{i=1}^n X_{i,j}X_{i,k}$. By the sub-Gaussian property, $\|X_{i,j}X_{i,k}\|_{\psi_1} \lesssim \|X_{i,j}\|_{\psi_2}\|X_{i,k}\|_{\psi_2} \lesssim 1$. Thus, by Bernstein's inequality, we have

$$\mathbb{P}\left\{|(\hat{\Sigma}_n)_{j,k} - (\Sigma)_{j,k}| \geqslant t\right\} \lesssim e^{-Cn\min(t,t^2)}.$$

Then, taking the union bound, we obtain

$$\mathbb{P}\left\{\sup_{j,k}|(\hat{\Sigma}_n)_{j,k} - \Sigma_{j,k}| \geqslant t\right\} \lesssim d^2 e^{-Cn\min(t,t^2)}.$$

Hence, for any $\delta \in (0, 1/e)$, w.p. $1-\delta$ the following holds for any $j,k \in [d]$

$$|(\hat{\Sigma}_n)_{j,k} - (\Sigma)_{j,k}| \lesssim \sqrt{\frac{\log(d/\delta)}{n}}.$$

□

D.1. Proof of Theorem 5.2

In this section, we prove Theorem 5.2 for the case of the spectral norm, Frobenius norm, and the trace separately.

Lemma D.2 (The spectral norm)**.** *Suppose Assumption 5.1 holds. For any $\delta \in (0,1)$, let $n \gtrsim d + \log(1/\delta)$ such that $\varepsilon_n = \sqrt{\frac{d+\log(1/\delta)}{n}} \leqslant 1$. Then, w.p. $1-\delta$, we have*

$$(1-\varepsilon_n)\|\theta\|_\infty \leqslant \|G(\theta)\|_2 \leqslant (1+\varepsilon_n)\|\theta\|_\infty.$$

Proof. Let $z = (u,v) \in \mathbb{R}^{2d}$ with $u,v \in \mathbb{R}^d$ such that $\|z\|^2 = \|u\|^2 + \|v\|^2 = 1$. Then, we have

$$\begin{aligned}
\|G(\theta)\|_2 &= \sup_{\|z\|=1} z^T G(\theta) z = \sup_{\|z\|=1} \sum_{j,k=1}^n (u_i u_j a_i a_j \hat{s}_{i,j} + u_i v_j a_i b_j \hat{s}_{i,j} + v_i u_j b_i a_j \hat{s}_{i,j} + v_i v_j b_i b_j \hat{s}_{i,j}) \\
&= \sup_{\|z\|=1} \Big((u\circ a)^T \hat{\Sigma}_n (u\circ a) + 2(u\circ a)^T \hat{\Sigma}_n (v\circ b) + (v\circ b)^T \hat{\Sigma}_n (v\circ b)\Big) \\
&= \sup_{\|z\|=1} (u\circ a + v\circ b)^T \hat{\Sigma}_n (u\circ a + v\circ b) \\
&= \sup_{\|z\|=1} \Big(\|u\circ a + v\circ b\|^2 + (u\circ a + v\circ b)^T (\hat{\Sigma}_n - I)(u\circ a + v\circ b)\Big).
\end{aligned}$$

By Lemma A.14, we have w.p. at least $1-\delta$ that $\|\hat{\Sigma}_n - I_d\| \leqslant \varepsilon_n$. Therefore,

$$(1-\varepsilon_n)\sup_{\|z\|=1}\|u\circ a + v\circ b\|^2 \leqslant \|G(\theta)\|_2 \leqslant (1+\varepsilon_n)\sup_{\|z\|=1}\|u\circ a + v\circ b\|^2.$$

Noticing that

$$\sup_{\|z\|=1}\|u\circ a + v\circ b\|^2 = \sup_{\|z\|=1}\sum_{j=1}^d (a_j u_j + b_j v_j)^2$$

$$
\begin{aligned}
&= \sup_{\|z\|=1} \sum_j (a_j^2 + b_j^2)(u_j^2 + v_j^2) \\
&= \sup_{t \in \mathbb{S}^{d-1}} \sum_j (a_j^2 + b_j^2) t_j^2 \\
&= \max_j (a_j^2 + b_j^2),
\end{aligned}
$$

we complete the proof. □

Lemma D.3. *For any $\delta \in (0,1)$, if $n \gtrsim \log(d/\delta)$, then w.p. at least $1-\delta$ that*

$$(1-\varepsilon_n)(\|a\|^2 + \|b\|^2) \leqslant \mathrm{Tr}(G(\theta)) \leqslant (1+\varepsilon_n)(\|a\|^2 + \|b\|^2).$$

Proof. Notice that $\mathrm{Tr}(G(\theta)) = \sum_{j=1}^d (a_i^2 \hat{s}_{i,i} + b_i^2 \hat{s}_{i,i})$. By Lemma D.1, we have w.p. $1-\delta$, for any $i \in [n]$, $|\hat{s}_{i,i} - 1| \leqslant \varepsilon_n$. Combining them completes the proof. □

Lemma D.4. *For any $\delta \in (0,1)$, if $n \gtrsim \log(d/\delta)$, then w.p. at least $1-\delta$ we have*

$$(1-\varepsilon_n)\|\alpha\|_2 \leqslant \|G(\theta)\|_F \leqslant \varepsilon_n \|\alpha\|_1 + \sqrt{1+4\varepsilon_n}\|\alpha\|_2$$

Proof. By the definition,

$$\|G(\theta)\|_F^2 = \sum_{j,k=1}^d \left(a_i^2 a_j^2 \hat{s}_{i,j}^2 + a_i^2 b_j^2 \hat{s}_{i,j} + b_i^2 a_j^2 \hat{s}_{i,j}^2 + b_i^2 b_j^2 \hat{s}_{i,j}^2\right) \tag{48}$$

By Lemma D.1, w.p. at least $1-\delta$ we have $|s_{i,j} - \hat{s}_{i,j}| \leqslant \varepsilon_n$ for any $i, j \in [n]$. Thus,

$$\hat{s}_{i,j}^2 = (\hat{s}_{i,j} - s_{i,j} + s_{i,j})^2 \in \begin{cases} [0, \varepsilon_n^2] & \text{if } i \neq j \\ [(1-\varepsilon_n)^2, (1+\varepsilon_n)^2] & \text{if } i = j \end{cases}.$$

Plugging it into (48) gives:

$$
\begin{aligned}
\|G(\theta)\|_F^2 &\leqslant \sum_{i \neq j} (a_i^2 a_j^2 + a_i^2 b_j^2 + b_i^2 a_j^2 + b_i^2 b_j^2)\varepsilon_n^2 + (1+\varepsilon_n)^2 \sum_i (a_i^4 + 2a_i^2 b_i^2 + b_i^4) \\
&\leqslant \sum_{i,j} (a_i^2 a_j^2 + a_i^2 b_j^2 + b_i^2 a_j^2 + b_i^2 b_j^2)\varepsilon_n^2 + ((1+\varepsilon_n)^2 + \varepsilon_n^2) \sum_i (a_i^4 + 2a_i^2 b_i^2 + b_i^4) \\
&\leqslant \varepsilon_n^2 (\sum_i a_i^2 + b_i^2)^2 + (1+2\varepsilon_n)^2 \sum_i (a_i^2 + b_i^2)^2,
\end{aligned} \tag{49}
$$

and

$$
\begin{aligned}
\|G(\theta)\|_F^2 &\geqslant \sum_i s_{i,i}^2 (a_i^4 + 2a_i^2 b_i^2 + b_i^4) \\
&\geqslant (1-\varepsilon_n)^2 \sum_i (a_i^2 + b_i^2)^2 = (1-\varepsilon_n)^2 \|\alpha\|_2^2.
\end{aligned} \tag{50}
$$

Combining (49) and (50) completes the proof. □

D.2. Proof of Theorem 5.3.

For any $Q > 0$, denote the class of linear predictors with bounded ℓ_1 by $\mathcal{H}_Q = \{h_\beta : x \to \beta^T x | \|\beta\|_1 \leqslant Q\}$, for which Shalev-Shwartz & Ben-David (2014, Lemma 26.11) gives

$$\mathfrak{R}_n(\mathcal{H}_Q) = \sup_{x_1,\ldots,x_n} \widehat{\mathrm{Rad}}_n(\mathcal{H}) \leqslant \max_{i \in [n]} \|x_i\|_\infty Q \sqrt{\frac{2\log(2d)}{n}} \leqslant Q\sqrt{\frac{2\log(2d)}{n}}.$$

Let $k = \|\beta_*\|_1$ and $\phi(z) = z^2/2$ be the loss function. Then, we have $|\phi''| \leqslant 1$ and $|\phi| \lesssim (Q+k)^2/2$ since for $\|\beta\|_1 \leqslant Q$:

$$\frac{1}{2}(\beta^T x - \beta_*^T x)^2 \leqslant \frac{1}{2}\|\beta - \beta_*\|_1^2 \|x\|_\infty \leqslant \frac{1}{2}(Q+k)^2.$$

Then, applying (15) to a minimizer $H \in \mathcal{H}_Q$ gives

$$\mathcal{R}(H) \lesssim \log^3(n) Q^2 \frac{\log(2d)}{n} + \frac{(Q+k)^2 \log(1/\delta)}{n}. \tag{51}$$

Now we turn to consider the linear predictor implemented by two-layer diagonal networks. Let $\hat{\theta} = (\hat{a}, \hat{b})$ and $\hat{\beta} = \hat{a} \odot \hat{b}$. By Proposition 3.2 and Theorem 5.2, we have w.p. $1-\delta$ that

$$\frac{2}{\eta} \geqslant \mathrm{Tr}(G(\hat{\theta})) \geqslant (1-\varepsilon_n) \sum_j (\hat{a}_j^2 + \hat{b}_j^2) \geqslant 2(1-\varepsilon_n)\|\hat{\beta}\|_1.$$

Therefore, $f(\cdot; \hat{\theta}) \in \mathcal{H}_{\hat{Q}}$ with $\hat{Q} \leqslant 1/(\eta(1-\varepsilon_n)) \lesssim 1/\eta$, where the last step is because that we assume n satisfies $\varepsilon_n \leqslant 1/2$. Plugging it into (51) gives

$$\mathcal{R}(\hat{\theta}) \leqslant \frac{(1/\eta)^2 \log^2(n) \log(d)}{n} + \frac{(k + \frac{1}{\eta})^2 \log(1/\delta)}{n}.$$

□