The Noise Geometry of Stochastic Gradient Descent: A Quantitative and Analytical Characterization

Anonymous Author(s)
Affiliation
Address
email

Abstract

Empirical studies have demonstrated that the noise in stochastic gradient descent (SGD) aligns favorably with the local geometry of loss landscape. However, theoretical and quantitative explanations for this phenomenon remain sparse. In this paper, we offer a comprehensive theoretical investigation into the aforementioned *noise geometry* for over-parameterized linear (OLMs) models and two-layer neural networks. We scrutinize both average and directional alignments, paying special attention to how factors like sample size and input data degeneracy affect the alignment strength. As a specific application, we leverage our noise geometry characterizations to study how SGD escapes from sharp minima, revealing that the escape direction has significant components along flat directions. This is in stark contrast to GD, which escapes only along the sharpest directions. To substantiate our theoretical findings, both synthetic and real-world experiments are provided.

1 Introduction

Stochastic gradient descent (SGD) and its variants have become the de facto optimizers for training machine learning models (Bottou, 1991). Unlike full-batch gradient descent (GD), SGD uses only mini-batches of data in each iteration, which injects noise into the optimization process. This noise can have a pronounced impact on both the convergence behavior (Thomas et al., 2020; Wojtowytsch, 2023; Feng and Tu, 2021; Simsekli et al., 2019) and the generalization capabilities (Zhang et al., 2017; Keskar et al., 2017; Wu et al., 2017; Zhu et al., 2019; Smith et al., 2020) of the algorithm.

Zhu et al. (2019); Wu et al. (2020); Xie et al. (2020) showed that SGD noise is highly anisotropic and in particular, the noise covariance matrix aligns well with the Hessian matrix. As such, they propose a Hessian-based approximation of the noise covariance: $\Sigma(\boldsymbol{\theta}) \approx \sigma^2 H(\boldsymbol{\theta})$, where $\Sigma(\boldsymbol{\theta})$ and $H(\boldsymbol{\theta})$ denote the noise covariance and Hessian matrices at $\boldsymbol{\theta}$, respectively and σ serves as a small constant denoting the noise magnitude. Subsequent works (Feng and Tu, 2021; Mori et al., 2022; Wojtowytsch, 2021; Liu et al., 2021) presented an improved Hessian-based approximation: $\Sigma(\boldsymbol{\theta}) \approx 2L(\boldsymbol{\theta})H(\boldsymbol{\theta})$ for regression problems with square loss, where $L(\boldsymbol{\theta})$ denotes the loss value. This refined approximation acknowledges the fact that the noise magnitude is proportional to the loss value.

However, the alignment between SGD noise and local landscape geometry remains empirical observations, lacking quantitative characterization and theoretical grounding. Hessian-based approximations are not accurate, as underscored by Thomas et al. (2020). A recent effort by Wu et al. (2022) employed a normalized cosine similarity between $\Sigma(\boldsymbol{\theta})$–which is close to the Hessian matrix in low loss regions–and the empirical Fisher matrix $G(\boldsymbol{\theta})$ as a metric to quantify the alignment. This metric is inspired by analyzing the dynamical stability of SGD (Wu et al., 2018) and can be interpreted as certain type of average alignment. Nevertheless, the analysis in Wu et al. (2022) is restricted to over-parameterized linear models (OLMs) and operates under the assumption of infinite data, leaving open questions about the generalizability of such alignment in more practically relevant settings.

Our contribution. Let n, d denote the sample size, input dimension, respectively. Then, our contributions can be summarized as follows.

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

- We first extend the average alignment analysis (Wu et al., 2022) to finite sample scenarios, offering a comprehensive investigation of how factors like sample size and input data degeneracy impact the alignment strength. We establish that, as long as $d_{\text{eff}} \gtrsim \log n$, the alignment strength is lower-bounded for both OLMs and two-layer neural networks–models not considered in Wu et al. (2022). Here, d_{eff} represents an effective input dimension, and this condition accommodates the important regimes like $n \sim \log(d_{\text{eff}})$ (for sparse recovery) and $n \sim d_{\text{eff}}$ (the proportional scaling).
- We then delve into a directional alignment analysis, probing whether the component of noise energy along a specific direction is proportional to the curvature in that direction. Our results show that for OLMs, as long as $n \gtrsim d$, the strength of directional alignment is lower-bounded acorss all directions and the entire parameter space.
- Lastly, we provide a detailed analysis of the mechanisms by which SGD escapes from sharp minima by leveraging our noise geometry results. We show that *the escape direction of SGD exhibits significant components along flat directions of the local landscape*. This stands in stark contrast to GD, which escapes from minima only along the sharpest direction. We also discuss the implications of this unique escape behavior, providing a preliminary explaination of how cyclical learning rate (Smith, 2017; Loshchilov and Hutter, 2017) can help find flatter minima.

It is worth noting that our theoretical guarantees apply effectively to both isotropic and anisotropic inputs, and *the guaranteed alignment strength is independent of the degree of overparameterization*. In addition, all theoretical findings are supported by numerical experiments conducted on both small-scale and larger-scale models, provided in Appendix C and D. Overall, our work advances the theoretical understanding of the geometry of SGD noise and provides insights into how SGD navigates the loss landscape.

2 Preliminaries

Let $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$ be the training set and $f(\cdot;\boldsymbol{\theta}) : \mathbb{R}^d \to \mathbb{R}$ be the model parameterized by $\boldsymbol{\theta} \in \mathbb{R}^p$. Let $\ell_i(\boldsymbol{\theta}) = \frac{1}{2}\left(f(\boldsymbol{x}_i;\boldsymbol{\theta}) - y_i\right)^2$ be the square loss at the i-th sample and $\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n \ell_i(\boldsymbol{\theta})$ be the empirical risk. To minimize $\mathcal{L}(\cdot)$, the mini-batch SGD updates as follows $\boldsymbol{\theta}(t+1) = \boldsymbol{\theta}(t) - \frac{\eta}{B}\sum_{i\in\mathcal{B}_t} \nabla \ell_i(\boldsymbol{\theta}(t))$, where $\mathcal{B}_t = \{\gamma_{t,1}, \cdots, \gamma_{t,B}\}$ is a batch with size $|\mathcal{B}_t| = B$, and $\gamma_{t,1}, \cdots, \gamma_{t,B} \overset{\text{i.i.d.}}{\sim} \mathbb{U}([n])$. To isolate the impact of noise, the SGD update is often reformulated as $\boldsymbol{\theta}(t+1) = \boldsymbol{\theta}(t) - \eta\left(\nabla\mathcal{L}(\boldsymbol{\theta}(t)) + \boldsymbol{\xi}(t)\right)$, where $\nabla\mathcal{L}(\boldsymbol{\theta}(t))$ is the full-batch gradient and $\boldsymbol{\xi}(t)$ represents the mini-batch noise satisfying $\mathbb{E}[\boldsymbol{\xi}(t)] = 0, \mathbb{E}[\boldsymbol{\xi}(t)\boldsymbol{\xi}(t)^\top] = \Sigma(\boldsymbol{\theta}(t))/B$ with the noise covariance given by $\Sigma(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n \nabla\ell_i(\boldsymbol{\theta})\nabla\ell_i(\boldsymbol{\theta})^\top - \nabla\mathcal{L}(\boldsymbol{\theta})\nabla\mathcal{L}(\boldsymbol{\theta})^\top$. In the above setup, the Hessian matrix of the empirical risk is given by $H(\boldsymbol{\theta}) = G(\boldsymbol{\theta}) + \frac{1}{n}\sum_{i=1}^n \left(f(\boldsymbol{x}_i;\boldsymbol{\theta}) - y_i\right)\nabla^2 f(\boldsymbol{x}_i;\boldsymbol{\theta})$, where $G(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n \nabla f(\boldsymbol{x}_i;\boldsymbol{\theta})\nabla f(\boldsymbol{x}_i;\boldsymbol{\theta})^\top$ is the empirical Fisher matrix. When the fit errors are small, we have $G(\boldsymbol{\theta}) \approx H(\boldsymbol{\theta})$ and in particular, for global minima $\boldsymbol{\theta}^*$, $H(\boldsymbol{\theta}^*) = G(\boldsymbol{\theta}^*)$. Additionally, for linear regression $f(\boldsymbol{x};\boldsymbol{\theta}) = \boldsymbol{\theta}^\top\boldsymbol{x}$, $H(\boldsymbol{\theta}) = G(\boldsymbol{\theta}) \equiv \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^\top$.

Over-parameterized linear models (OLMs). An OLM is defined as $f(\boldsymbol{x};\boldsymbol{\theta}) = F(\boldsymbol{\theta})^\top\boldsymbol{x}$, where $F : \mathbb{R}^p \to \mathbb{R}^d$ denotes a general re-parameterization function. Although $f(\cdot;\boldsymbol{\theta})$ only represents linear functions, the corresponding loss landscape can be highly non-convex. Some typical examples include (i) the linear model $F(\boldsymbol{w}) = \boldsymbol{w}$; (ii) the diagonal linear network: $F(\boldsymbol{\theta}) = (\alpha_1^2 - \beta_1^2, \dots, \alpha_d^2 - \beta_d^2)^\top$; and (iii) the linear network: $F(\boldsymbol{\theta}) = W_1 W_2 \cdots W_L$. Notably, OLMs have been widely used to analyze the optimization and implicit bias of SGD (Arora et al., 2019; Woodworth et al., 2020; Pesme et al., 2021; HaoChen et al., 2021; Azulay et al., 2021).

Noise Geometry. Before proceeding to our refined characterization of the noise geometry, we first recall two existing results on quantifying the geometry of SGD noise.

- Mori et al. (2022) proposed the following Hessian-based approximation: $\Sigma(\boldsymbol{\theta}) \approx 2\mathcal{L}(\boldsymbol{\theta})G(\boldsymbol{\theta})$. It reveals 1) the noise magnitude is proportional to the loss value; 2) the noise covariance aligns with the Fisher matrix. This approximation is intuitive and helpful for understanding, but it cannot be accurate in general.
- *Online SGD for OLMs with Gaussian inputs*. Suppose $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, S)$ and $n = \infty$ (i.e., online SGD). For OLMs, Wu et al. (2022) derived the following analytical expression $\Sigma(\boldsymbol{\theta}) = 2\mathcal{L}(\boldsymbol{\theta})G(\boldsymbol{\theta}) + \nabla\mathcal{L}(\boldsymbol{\theta})\nabla\mathcal{L}(\boldsymbol{\theta})^\top$.

3 Average Alignment

Let $\Sigma_1(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n \nabla\ell_i(\boldsymbol{\theta})\nabla\ell_i(\boldsymbol{\theta})^\top, \Sigma_2(\theta) = \nabla\mathcal{L}(\boldsymbol{\theta})\nabla\mathcal{L}(\boldsymbol{\theta})^\top$. Then $\Sigma(\boldsymbol{\theta}) = \Sigma_1(\boldsymbol{\theta}) - \Sigma_2(\boldsymbol{\theta})$. It is commonly believed that the magnitude of the full-batch gradient $\nabla\mathcal{L}$ is relatively small compared to the sample gradients $\{\nabla\ell_i\}_i$. Consequently, the influence of $\Sigma_2(\boldsymbol{\theta})$ would be negligible compared to $\Sigma_1(\boldsymbol{\theta})$. Following Wu et al. (2022), we consider the following metrics of quantifying average alignment: $\mu(\boldsymbol{\theta}) = \frac{\mathrm{Tr}(\Sigma(\boldsymbol{\theta})G(\boldsymbol{\theta}))}{2\mathcal{L}(\boldsymbol{\theta})\|G(\boldsymbol{\theta})\|_{\mathrm{F}}^2}$.

Wu et al. (2022) guarantees $\mu(\boldsymbol{\theta}) \geq 1$ in an infinite data scenario. The following theorem extends it to finite-sample cases and the proof can be found in Appendix G. To simplify the statement, we define the effective dimension of inputs as $d_{\text{eff}} := \min\{\mathrm{srk}(S), \mathrm{srk}(S^2)\}$, where S represents the input covariance matrix and $\mathrm{srk}(S) = \mathrm{tr}(S)/\|S\|_2$ is the stable rank of S. In particular, when S is isotropic, we have $d_{\text{eff}} = d$.

Theorem 3.1 (OLM)**.** *Consider OLMs and assume $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, S)$. For any $\epsilon, \delta \in (0,1)$,*

(a) if $n/\log(n/\delta) \gtrsim 1/\epsilon^2$ and $d_{\text{eff}} \gtrsim \log(n/\delta)/\epsilon^2$, then w.p. *at least $1-\delta$, it holds that $\inf_{\boldsymbol{\theta}\in\mathbb{R}^p} \mu(\boldsymbol{\theta}) \geq \frac{(1-\epsilon)^2}{(1+\epsilon)^2 \mathrm{cond}^2(\nabla F(\boldsymbol{\theta})\nabla F(\boldsymbol{\theta})^\top)}$;*

(b) if $n \gtrsim d + \log(1/\delta)$, then w.p. *at least $1-\delta$, it holds that $\inf_{\boldsymbol{\theta}\in\mathbb{R}^p} \mu(\boldsymbol{\theta}) \gtrsim 1$.*

Result (a) is established by leveraging the high dimensionality of inputs, as stated by the condition $d_{\text{eff}} \gtrsim \log n$, which is particularly relevant for low-sample regimes. Notably, this includes the important regimes like $n \sim \log(d_{\text{eff}})$ (for sparse recovery) and $n \sim d_{\text{eff}}$ (the proportional scaling). In contrast, result (b) is pertinent to the enough-data regime where $n \gtrsim d$. Notably, the alignment holds no matter how degenerate the covariance matrix is. In a summary, these two results are complementary and collectively span all the regimes of interest.

Example. Consider the isotropic case where $S = I_d$ and linear regression $F(\boldsymbol{w}) = \boldsymbol{w}$. In this case, $\nabla F(\boldsymbol{w}) \equiv I_d$ and thus, Theorem 3.1 implies that it holds that $\inf_{\boldsymbol{\theta}\in\mathbb{R}^p} \mu(\boldsymbol{\theta}) \gtrsim 1$ as long as $n \gtrsim 1$.

Consider two-layer neural networks given by $f(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{k=1}^m a_k \phi(\boldsymbol{b}_k^\top \boldsymbol{x})$ with $a_k \in \{\pm 1\}$ to be fixed. We use $\boldsymbol{\theta} = (\boldsymbol{b}_1^\top, \cdots, \boldsymbol{b}_m^\top)^\top \in \mathbb{R}^{md}$ to denote the concatenation of all trainable parameters. Here, $\phi : \mathbb{R} \mapsto \mathbb{R}$ is an activation function with a nondegenerate derivative as defined below.

Assumption 3.2. There exist constants $\beta > \alpha > 0$ such that $\alpha \leq \phi'(z) \leq \beta$ holds for any $z \in \mathbb{R}$.

Example 3.3. *(i) A typical activation function that satisfies Assumption 3.2 is α-Leacky ReLU: $\phi(z) = \max\{\alpha z, z\}$, where $\alpha \in (0,1)$. (ii) Moreover, the assumption also holds for Sigmoid with the truncation trick (to prevent gradient vanishing of Sigmoid): $\phi(z) = 1/(1+\exp(-\mathrm{sgn}(z)\min\{|z|, M\}))$, where $M > 0$ is the truncation constant.*

Theorem 3.4 (2NN)**.** *Consider the two-layer network $f(\cdot;\boldsymbol{\theta})$ with the activation function satisfying Assumption 3.2 and assume $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, S)$. For any $\epsilon, \delta \in (0,1)$, if $n/\log(n/\delta) \gtrsim 1/\epsilon^2$ and $d_{\text{eff}} \gtrsim \log(n/\delta)/\epsilon^2$, then* w.p. *at least $1-\delta$, it holds that $\inf\limits_{\boldsymbol{\theta}\in\mathbb{R}^{md}} \mu(\boldsymbol{\theta}) \geq \frac{\alpha^2(1-\epsilon)^2}{\beta^2(1+\epsilon)^2}$.*

Remark 3.5. We would like to emphasize that the conditions presented in Theorem 3.1 and 3.4 are independent of the model size p.

The numerical validation is referred to Appendix C.

4 Directional Alignment

In Section 3, we focused solely on average alignment. Subsequently, we delve into a specific type of directional alignment: *whether noise energy along a direction is proportional to the curvature of loss landscape along that direction*. To this end, we define the following metric to measure the strength of directional alignment.

Definition 4.1 (Directional Alignment)**.** Given $\boldsymbol{v} \in \mathbb{R}^p$, the alignment along $\boldsymbol{v}$ is defined as $g(\boldsymbol{\theta};\boldsymbol{v}) := \frac{\boldsymbol{v}^\top\Sigma(\boldsymbol{\theta})\boldsymbol{v}}{2\mathcal{L}(\boldsymbol{\theta})(\boldsymbol{v}^\top G(\boldsymbol{\theta})\boldsymbol{v})}$, where $\boldsymbol{v}^\top\Sigma(\boldsymbol{\theta})\boldsymbol{v} = \mathbb{E}[(\boldsymbol{\xi}(\boldsymbol{\theta})^\top\boldsymbol{v})^2]$ denotes the noise energy along direction $\boldsymbol{v}$, $\boldsymbol{v}^\top G(\boldsymbol{\theta})\boldsymbol{v}$ is the curvature of loss landscape along $\boldsymbol{v}$, and $2\mathcal{L}(\boldsymbol{\theta})$ is only a scaling factor.

Theorem 4.2 (One-sided bound)**.** *Consider OLMs and assume $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, S)$. For any $\delta \in (0,1)$, if $n \gtrsim d + \log(1/\delta)$, then* w.p. *at least $1-\delta$, we have $\inf_{\boldsymbol{\theta},\boldsymbol{v}\in\mathbb{R}^p} g(\boldsymbol{\theta};\boldsymbol{v}) \gtrsim 1$.*

This theorem establishes that a sample size satisfying $n \gtrsim d$ is sufficient to guarantee a uniform lower bound for alignment across all directions and the entire parameter space. The subsequent theorem builds upon this by offering a two-sided bound on alignment strength, albeit at the cost of requiring a larger sample size.

Theorem 4.3 (Two-sided bound)**.** *Consider OLMs and assume* $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, S)$. *For any* $\epsilon, \delta \in (0,1)$, *if* $n \gtrsim \max\left\{\left(d^2 \log^2(1/\epsilon) + \log^2(1/\delta)\right)/\epsilon, \left(d \log(1/\epsilon) + \log(1/\delta)\right)/\epsilon^2\right\}$, *then* w.p. *at least* $1-\delta$, *we have the following two-side uniform bounds for the directional alignment: (i).* $\frac{1-\epsilon}{(1+\epsilon)^2} \leq \inf_{\boldsymbol{\theta}, \boldsymbol{v} \in \mathbb{R}^p} g(\boldsymbol{\theta}; \boldsymbol{v}) \leq \sup_{\boldsymbol{\theta}, \boldsymbol{v} \in \mathbb{R}^p} g(\boldsymbol{\theta}; \boldsymbol{v}) \leq \frac{2+\epsilon}{(1-\epsilon)^2}$; *(ii).* $\frac{1-\epsilon}{(1+\epsilon)^2} \leq \inf_{\boldsymbol{\theta} \in \mathbb{R}^p, \langle \boldsymbol{v}, \nabla \mathcal{L}(\boldsymbol{\theta}) \rangle = 0} g(\boldsymbol{\theta}; \boldsymbol{v}) \leq \sup_{\boldsymbol{\theta} \in \mathbb{R}^p, \langle \boldsymbol{v}, \nabla \mathcal{L}(\boldsymbol{\theta}) \rangle = 0} g(\boldsymbol{\theta}; \boldsymbol{v}) \leq \frac{1+\epsilon}{(1-\epsilon)^2}$.

Notably, for directions satisfying $\boldsymbol{v} \perp \nabla \mathcal{L}(\boldsymbol{\theta})$, the alignment strength is nearly 1. The proofs of the above two theorems are deferred to Appendix H. The numerical validation is referred to Appendix C.

5 How SGD Escapes from Sharp Minima

Let $\boldsymbol{w} = \boldsymbol{\theta} - \boldsymbol{\theta}^*$ and $\boldsymbol{z}_i = \nabla f(\boldsymbol{x}_i; \boldsymbol{\theta}^*)$. Then, $G(\boldsymbol{\theta}^*) = \frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i \boldsymbol{z}_i^\top$ and the linearized SGD of iterates as follows $\boldsymbol{w}(t+1) = \boldsymbol{w}(t) - \eta\left(G(\boldsymbol{\theta}^*)\boldsymbol{w}(t) + \boldsymbol{\xi}(t)\right)$, where $\boldsymbol{\xi}(t)$ is the SGD noise. In addition, in this section, we simply use $\mathcal{L}(\boldsymbol{w}) = \frac{1}{2}\boldsymbol{w}^T G(\boldsymbol{\theta}^*)\boldsymbol{w}$ to denote the corresponding loss. We make the following assumption on the noise alignment.

Assumption 5.1 (Eigen-directional alignment)**.** let $G(\boldsymbol{\theta}^*) = \sum_{i=1}^d \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^\top$ be the eigen decomposition of $G(\boldsymbol{\theta}^*)$. Assume that there exist $A_1, A_2 > 0$ such that it holds for any $\boldsymbol{w} \in \mathbb{R}^d$, $A_1 \mathcal{L}(\boldsymbol{w})\lambda_i \leq \mathbb{E}[|\xi(\boldsymbol{w})^\top \boldsymbol{u}_i|^2] \leq A_2 \mathcal{L}(\boldsymbol{w})\lambda_i$.

For linear models under the setting of Theorem 4.3, Assumption 5.1 is provably valid. It is important to clarify, however, that the above assumption only requires the alingment along eigen-directions, which is considerably less stringent compared to the uniform directional alignment specified in Theorem 4.3.

Eigen-decomposition of SGD. By leveraging Assumption 5.1, we can analyze the SGD dynamics in the eigenspace. Let $\boldsymbol{w}(t) = \sum_{i=1}^d w_i(t)\boldsymbol{u}_i$ with $w_i(t) = \boldsymbol{u}_i^\top \boldsymbol{w}(t)$. Then, $w_i(t+1) = (1 - \eta\lambda_i)w_i(t) + \eta\boldsymbol{\xi}(t)^\top \boldsymbol{u}_i$. Taking the expectation of the square of both sides, we obtain $\mathbb{E}\left[w_i^2(t+1)\right] = (1-\eta\lambda_i)^2 \mathbb{E}\left[w_i^2(t)\right] + \eta^2 \mathbb{E}[|\boldsymbol{u}_i^\top \boldsymbol{\xi}(t)|^2]$, where the noise term: $\mathbb{E}[|\boldsymbol{u}_i^\top \boldsymbol{\xi}(t)|^2] \sim \lambda_i \mathcal{L}(\boldsymbol{w}_t)$ according to Assumption 5.1.

Let $X_t = \sum_{i=1}^k \lambda_i \mathbb{E}[w_i^2(t)], Y_t = \sum_{i=k+1}^d \lambda_i \mathbb{E}[w_i^2(t)]$, denoting the components of loss energy along sharp and flat directions, respectively. Let $D_k(t) = Y_t / X_t$, which measures the concentration of loss energy along flat directions. Analogously, let $P_k(t) = \sum_{i=k+1}^d \mathbb{E}[w_i^2(t)] / \sum_{i=1}^k \mathbb{E}[w_i^2(t)]$, which measure the concentration of variance along flat directions. It is easy to show that $P_k(t) \geq D_k(t)\lambda_k/\lambda_{k+1}$. Therefore, when λ_k/λ_{k+1} is lower bounded, a concentration of loss energy along flat directions can lead to a similar concentration in terms of variance.

Theorem 5.2 (Escape of SGD)**.** *Suppose Assumption 5.1 holds and let* $\eta = \frac{\beta}{\|G(\boldsymbol{\theta}^*)\|_F}$. *Then, there exists absolute constants* $c_1, c_2 > 0$ *such that if* $\beta \geq c_1$, *then SGD will escape from that minima and for any* $k \in [d]$, *it holds that when* $t \geq \max\left\{1, \frac{\log\left(c_2/\eta(\sum_{i=1}^k \lambda_i^2)^{1/2}\right)}{\log \beta}\right\}$*:* $D_k(t) \gtrsim \frac{\sum_{i=k+1}^d \lambda_i^2}{\sum_{i=1}^k \lambda_i^2}$.

The proof can be found in Appendix I. This theorem reveals that during SGD's escape process, the loss rapidly accumulates a significant component along flat directions of the loss landscape. The precise loss ratio between the flat and sharp directions is governed by the spectrum of Hessian matrix. In particular, $D_1(t) \gtrsim \operatorname{srk}(G^2) - 1$, indicating that in high dimension, i.e., $\operatorname{srk}(G^2) \gg 1$, the loss energy along the sharpest directions becomes negligible during the SGD's escape process. This stands in stark contrast to GD, which always escapes along the sharpest direction:

Proposition 5.3 (Escape of GD)**.** *Consider GD with learning rate* $\eta = \beta/\lambda_1$. *If* $\beta > 2$, *then* $D_1(t) \leq \sum_{i=2}^d \frac{\lambda_i(1-\eta\lambda_i)^{2t} w_i^2(0)}{\lambda_1(1-\eta\lambda_1)^{2t} w_1^2(0)}$.

In particular, if $w_1(0) \neq 0$ and $\lambda_1 > \lambda_2$, then the above proposition implies that $D_1(t)$ decreases to 0 exponentially fast for GD. The numerical validation is referred to Appendix C.

Furthermore, as an implication of SGD's escaping direction, we explain the **implicit bias** of **cyclical learning rate** in Appendix B.

References

Kwangjun Ahn, Jingzhao Zhang, and Suvrit Sra. Understanding the unstable convergence of gradient descent. In *International Conference on Machine Learning*, pages 247–257. PMLR, 2022.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems*, 32, 2019.

Shahar Azulay, Edward Moroshko, Mor Shpigel Nacson, Blake E Woodworth, Nathan Srebro, Amir Globerson, and Daniel Soudry. On the implicit bias of initialization shape: Beyond infinitesimal mirror descent. In *International Conference on Machine Learning*, pages 468–477. PMLR, 2021.

Léon Bottou. Stochastic gradient learning in neural networks. *Proceedings of Neuro-Nımes*, 91(8), 1991.

Jian-Feng Cai, Meng Huang, Dong Li, and Yang Wang. Nearly optimal bounds for the global geometric landscape of phase retrieval. *arXiv preprint arXiv:2204.09416*, 2022.

Jeremy Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2020.

Hadi Daneshmand, Jonas Kohler, Aurelien Lucchi, and Thomas Hofmann. Escaping saddles with stochastic gradients. In *International Conference on Machine Learning*, pages 1155–1164. PMLR, 2018.

Yu Feng and Yuhai Tu. 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.

Jonas Geiping, Micah Goldblum, Phillip E Pope, Michael Moeller, and Tom Goldstein. Stochastic training is not necessary for generalization. *arXiv preprint arXiv:2109.14119*, 2021.

Botao Hao, Yasin Abbasi Yadkori, Zheng Wen, and Guang Cheng. Bootstrapping upper confidence bound. *Advances in neural information processing systems*, 32, 2019.

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

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

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

Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E Hopcroft, and Kilian Q Weinberger. Snapshot ensembles: Train 1, get M for free. In *International Conference on Learning Representations*, 2018.

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

Bobby Kleinberg, Yuanzhi Li, and Yang Yuan. An alternative view: When does SGD escape local minima? In *International Conference on Machine Learning*, pages 2698–2707. PMLR, 2018.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images, 2009. URL https://www.cs.toronto.edu/~kriz/cifar.html.

Sungyoon Lee and Cheongjae Jang. A new characterization of the edge of stability based on a sharpness measure aware of batch gradient distribution. In *The Eleventh International Conference on Learning Representations*, 2022.

Kangqiao Liu, Liu Ziyin, and Masahito Ueda. Noise and fluctuation of finite learning rate stochastic gradient descent. In *International Conference on Machine Learning*, pages 7045–7056. PMLR, 2021.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017.

Chao Ma, Daniel Kunin, Lei Wu, and Lexing Ying. Beyond the quadratic approximation: The multiscale structure of neural network loss landscapes. *Journal of Machine Learning*, 1(3): 247–267, 2022.

Takashi Mori, Liu Ziyin, Kangqiao Liu, and Masahito Ueda. Power-law escape rate of SGD. In *International Conference on Machine Learning*, pages 15959–15975. PMLR, 2022.

Scott Pesme, Loucas Pillaud-Vivien, and Nicolas Flammarion. Implicit bias of SGD for diagonal linear networks: a provable benefit of stochasticity. *Advances in Neural Information Processing Systems*, 34, 2021.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015*, 2015.

Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, pages 5827–5837. PMLR, 2019.

Leslie N Smith. Cyclical learning rates for training neural networks. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pages 464–472. IEEE, 2017.

Samuel Smith, Erich Elsen, and Soham De. On the generalization benefit of noise in stochastic gradient descent. In *International Conference on Machine Learning*, pages 9058–9067. PMLR, 2020.

Valentin Thomas, Fabian Pedregosa, Bart Merriënboer, Pierre-Antoine Manzagol, Yoshua Bengio, and Nicolas Le Roux. On the interplay between noise and curvature and its effect on optimization and generalization. In *International Conference on Artificial Intelligence and Statistics*, pages 3503–3513. PMLR, 2020.

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

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

Stephan Wojtowytsch. Stochastic gradient descent with noise of machine learning type part i: Discrete time analysis. *Journal of Nonlinear Science*, 33(3):45, 2023.

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

Jingfeng Wu, Wenqing Hu, Haoyi Xiong, Jun Huan, Vladimir Braverman, and Zhanxing Zhu. On the noisy gradient descent that generalizes as SGD. In *International Conference on Machine Learning*, pages 10367–10376. PMLR, 2020.

Lei Wu and Weijie J Su. The implicit regularization of dynamical stability in stochastic gradient descent. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 37656–37684. PMLR, 23–29 Jul 2023.

Lei Wu, Zhanxing Zhu, and Weinan E. Towards understanding generalization of deep learning: Perspective of loss landscapes. *arXiv preprint arXiv:1706.10239*, 2017.

Lei Wu, Chao Ma, and Weinan E. How SGD selects the global minima in over-parameterized learning: A dynamical stability perspective. *Advances in Neural Information Processing Systems*, 31:8279–8288, 2018.

Lei Wu, Mingze Wang, and Weijie J Su. The alignment property of SGD noise and how it helps select flat minima: A stability analysis. *Advances in Neural Information Processing Systems*, 35: 4680–4693, 2022.

Zeke Xie, Issei Sato, and Masashi Sugiyama. A diffusion theory for deep learning dynamics: Stochastic gradient descent exponentially favors flat minima. In *International Conference on Learning Representations*, 2020.

Zeke Xie, Xinrui Wang, Huishuai Zhang, Issei Sato, and Masashi Sugiyama. Adaptive inertia: Disentangling the effects of adaptive learning rate and momentum. In *International conference on machine learning*, pages 24430–24459. PMLR, 2022.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

Hongyi Zhang, Yann N. Dauphin, and Tengyu Ma. Residual learning without normalization via better initialization. In *International Conference on Learning Representations*, 2019.

Pan Zhou, Jiashi Feng, Chao Ma, Caiming Xiong, Steven Chu Hong Hoi, and Weinan E. Towards theoretically understanding why SGD generalizes better than Adam in deep learning. *Advances in Neural Information Processing Systems*, 33, 2020.

Zhanxing Zhu, Jingfeng Wu, Bing Yu, Lei Wu, and Jinwen Ma. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects. In *International Conference on Machine Learning*, pages 7654–7663. PMLR, 2019.

Liu Ziyin, Kangqiao Liu, Takashi Mori, and Masahito Ueda. Strength of minibatch noise in SGD. In *International Conference on Learning Representations*, 2022.

Appendix

A Other Related Work

Noise geometry. Ziyin et al. (2022) provides a detailed analysis of the noise structure of online SGD for linear regression. We instead consider nonlinear models and finite-sample regimes. We also acknowledge the existence of works such as Simsekli et al. (2019); Zhou et al. (2020), which argue that the magnitude of SGD noise is heavy-tailed. However, our particular focus is on the noise shape and the observation that the noise magnitude is directly proportional to the loss value.

Escape from minima and saddle points The phenomenon of SGD escaping from sharp minima exponentially fast was initially studied in Zhu et al. (2019) as an indicator of how much SGD dislikes sharp minima. This provides an explanation of the famous "flat minima hypothesis" (Hochreiter and Schmidhuber, 1997; Keskar et al., 2017; Wu and Su, 2023)—one of the most important observations in explaining the implicit regularization of SGD. However, existing analyses of the escape phenomenon have primarily focused on the escape rate (Wu et al., 2018; Zhu et al., 2019; Xie et al., 2020; Mori et al., 2022; Ziyin et al., 2022). In contrast, we extends this focus by providing analysis of escape direction, which is enabled by our characterizations of the noise geometry. Kleinberg et al. (2018) introduced an alternative perspective, positing that SGD circumvents local minima by navigating an effective loss landscape that results from the convolution of the original landscape with SGD noise. In this context, our noise geometry characterizations can be beneficial in understanding the effective loss landscape. In addition, prior works like (Daneshmand et al., 2018; Xie et al., 2022) has illustrated that the alignment of noise with local geometry facilitates the rapid saddle-point escape of SGD. Our work offers theoretical substantiation for the alignment assumptions in these studies.

B Explaining the Implicit Bias of Cyclical Learning Rate

Gaining insights into the escape direction of SGD can be valuable for understanding its optimization dynamics, generalization properties, and the overall behavior. A more detailed discussion on this topic is available in Section E. In this section, however, we concentrate a specific example, illustrating the role of escape direction in enhancing the implicit bias of SGD through Cyclical Learning Rate (CLR) (Smith, 2017; Loshchilov and Hutter, 2017). As shown in Figure 2 of Huang et al. (2018), utilizing CLR enables SGD to cyclically escapes from (when increasing LR) and slides into (when decreasing LR) sharp regions, ultimately progressing towards flatter minima. We hypothesize that escape along flat directions plays a pivotal role in guiding SGD towards flatter region in this process.

Following Ma et al. (2022), we consider a toy OLM $f(x; \boldsymbol{w}) = (w_2/\sqrt{w_1^2+1})x$ with $x \sim \mathcal{N}(0,1)$. For simplicity, we consider the online setting, where the landscape

$$\mathcal{L}(\boldsymbol{w}) = w_2^2/[2(w_1^2+1)].$$

The global minima valley is $S = \{\boldsymbol{w} : w_2 = 0\}$ and for $\boldsymbol{w} \in S$, $\mathrm{tr}[\nabla^2 \mathcal{L}(\boldsymbol{w})] = 1/(1+w_1^2)$. Hence, the minimum gets flatter along the valley S when $|w_1|$ grows up. In Figure 1, we visualize the trajectories for both SGD+CLR and GD+CLR. One can observe that

- SGD escape from the minima along both the flat direction $\boldsymbol{e}_1$ and sharp direction $\boldsymbol{e}_2$. The component of along $\boldsymbol{e}_1$ leads to considerable increase in $w_1^2(t)$, facilitating the movement towards flatter region along the minimum valley S.
- On the contrary, GD escapes only along $\boldsymbol{e}_2$, yielding no increase in $w_1^2(t)$. Thus, we cannot observe clear movement towards flatter region for GD+CLR.

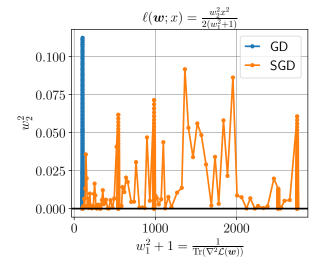


Figure 1: Visualization of the trajectories of SGD+CLR v.s. GD+CLR for our toy model. Both cases use the same CLR schedule. We can observe that SGD+CLR moves significantly towards flatter region, while GD+CLR only osscilates along the sharpest direction. We have extensively tuned the learning rates for GD+CLR but do not obseve significant movement towards flatter region in any case.

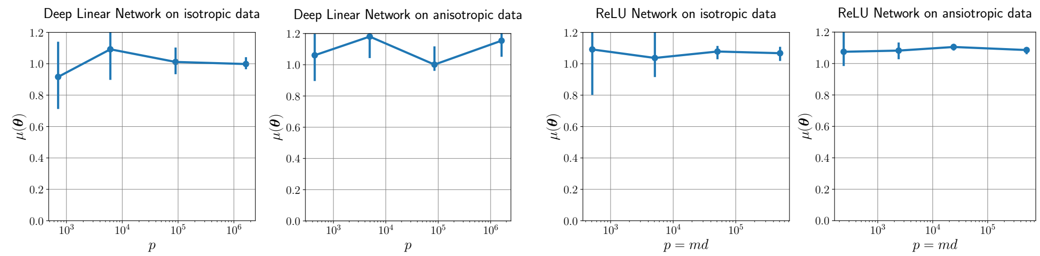
Thus, in this toy model, the fact that SGD escapes along flat directions is crucial in amplifying the implicit bias towards flat minima.

Nonetheless, understanding how the above mechanism manifests in practice remains an open question that warrants further investigation. We defer this topic to future work, as the primary focus of this paper is to understand the noise geometry rather than exhaustively explore its applications.

C Small-scale Experiments

C.1 Average Alignment

In this section, we present small-scale experiments to corroborate our theoretical results with a 4-layer linear network and two-layer ReLU network (both layers are trainable). Both isotropic and anistropic input distributions are examined and in paritucular, for the anistropic case, we set $\lambda_k^2(S) = 1/\sqrt{k}$. As for sample size, we set $n = 5\log(d_{\text{eff}})$ to focus on the low-sample regime. The results are reported in Figure 2 and it is evident that across all examined scenarios, the alignment strength is consistently lower-bounded and independent of the model size.



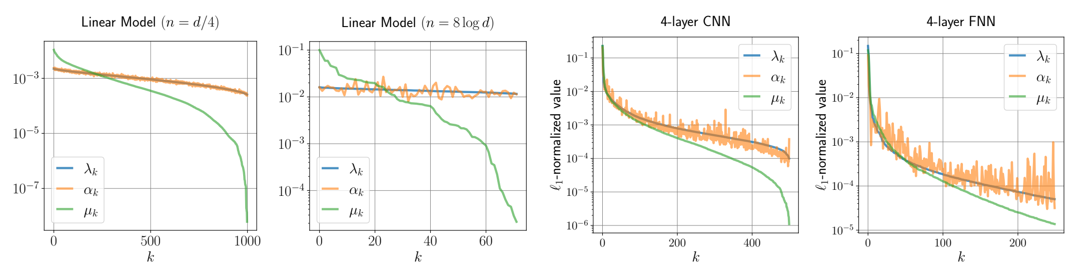
(a) Linear networks (b) Two-layer ReLU networks

Figure 2: The alignment strength is independent of model size. Two types of models: 4-layer linear network, and two-layer neural network are examined. In experiments, we set $n = 5\log(d_{\text{eff}}), d_{\text{eff}} = 50$. The error bar corresponds to the standard deviation over 20 independent runs.

C.2 Directional Alignment

In this experiment, we consider the alignment along the eigen-directions of Hessian matrix. Let $G(\boldsymbol{\theta}) = \sum_k \lambda_k(\boldsymbol{\theta})\boldsymbol{u}_k(\boldsymbol{\theta})\boldsymbol{u}_k(\boldsymbol{\theta})^\top$ be the eigen-decomposition of $G(\boldsymbol{\theta})$ respectively, where $\{\lambda_k(\boldsymbol{\theta})\}_k$ are the eigenvalues in a decreasing order and $\{\boldsymbol{u}_k(\boldsymbol{\theta})\}$ are the corresponding eigen-directions. Note that $\lambda_k(\boldsymbol{\theta})$ is the curvature of local landscape along $\boldsymbol{u}_k(\boldsymbol{\theta})$. Decompose SGD noise along these eigen-directions: $\boldsymbol{\xi}(\boldsymbol{\theta}) = \sum_k r_k(\boldsymbol{\theta})\boldsymbol{u}_k(\boldsymbol{\theta})$, where $r_k(\boldsymbol{\theta}) = \boldsymbol{\xi}(\boldsymbol{\theta})^\top\boldsymbol{u}_k(\boldsymbol{\theta})$ denotes the noise component in the direction of $\boldsymbol{u}_k(\boldsymbol{\theta})$. Consequently, the (scaled) expected noise magnitude in the direction $\boldsymbol{u}_k(\boldsymbol{\theta})$ is given by $\alpha_k(\boldsymbol{\theta}) = \mathbb{E}[r_k^2(\boldsymbol{\theta})]/2\mathcal{L}(\boldsymbol{\theta}) = \boldsymbol{u}_k^\top\Sigma(\boldsymbol{\theta})\boldsymbol{u}_k(\boldsymbol{\theta})/2\mathcal{L}(\boldsymbol{\theta})$. For comparison, let $\{\mu_k(\boldsymbol{\theta})\}_k$ denote the eigenvalues of $\Sigma(\boldsymbol{\theta})/2\mathcal{L}(\boldsymbol{\theta})$. When clear from the context, we will omit dependence on $\boldsymbol{\theta}$ for simplicity.

In Figure 3a, we examine linear regression in the regimes with limited data. Surprisingly, even with significantly fewer samples, we still observed that the noise energy along each eigen-direction remained roughly proportional to the corresponding curvature and the ratio is close 1. However, we noticed that the eigenvalues of $\Sigma(\boldsymbol{\theta})/2\mathcal{L}(\boldsymbol{\theta})$ decayed much faster than that of $G(\boldsymbol{\theta})$, indicating that the condition $n \gtrsim d$ stated in Theorem 4.2 is necessary to ensure uniform alignment across all directions. In Figure 3b, we further consider the classification of CIFAR-10 with a small convolutional neural network (CNN) and fully-connected neural network (FNN). We can see that the obsevation is consistent with Figure 3a, where the alignment along eigen-directions is significant.



(a) Linear models (b) Small neural networks

Figure 3: How the components of noise energy in *eigen-directions* $\{\alpha_k\}_k$ are proportional to the corresponding curvatures $\{\lambda_k\}_k$. α_k/λ_k can reflect the directional alignment along the eigen-directions of the local landscape. The eigenvalues of $\Sigma/2\mathcal{L}$ are also plotted as comparison. (a) Linear models on Gaussian data in the regimes with limited data, where we fix $d = 10^3$ and change n accordingly $(n = d/4, n = 8\log d)$. (b) 4-layer CNN and 4-layer FNN on CIFAR-10 dataset. For more experimental details, we refer to Appendix F.

C.3 Escaping Direction

Figure 4 presents numerical comparisons of the escaping directions between SGD and GD. It is evident that $D_1(t)$ exponentially decreases to zero for GD, indicating that GD escapes along the sharpest direction. In contrast, for SGD, $D_1(t)$ remains significantly large, indicating that SGD retains a substantial component along the flat directions during the escape process. Furhermore, the value of $D_1(t)$ positively correlates with $\text{srk}(G^2)$, as predicted by our Theorem 5.2. These observations provide empirical confirmation of our theoretical predictions.

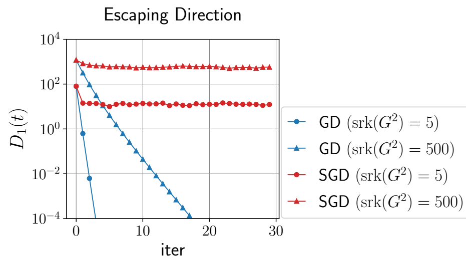
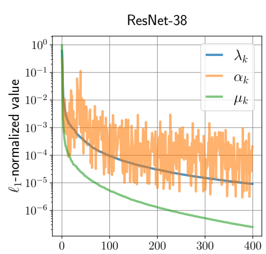


Figure 4: Comparison of escape directions between SGD and GD. The problem is linear regression and both SGD and GD are initialized near the global minimum by $\boldsymbol{w}(0) \sim \mathcal{N}(\boldsymbol{w}^*, e^{-10} I_d/d)$. To ensure escape, we choose $\eta = 1.2/\|G\|_{\mathrm{F}}$ and $\eta = 4/(\lambda_1 + \lambda_2)$ for SGD and GD, respectively. Please refer to Appendix F for more experimental details.

D Larger-scale Experiments for Deep Neural Networks

We have already provided small-scale experiments to confirm our theoretical findings. We now turn to justify the practical relevance by examining the classification of CIFAR-10 dataset (Krizhevsky and Hinton, 2009) with practical VGG nets (Simonyan and Zisserman, 2015) and ResNets (He et al., 2016). Note that larger-scale experiments on average alignment have been previously presented in Wu et al. (2022). Thus, our focus here is on investigate the directional alignment and escape direction of SGD. We refer to Appendix F for experimental details.

The directional alignment along eigen-directions. Figure 5 presents the directional alignments of SGD noise for ResNet-38 and VGG-13. The alignment is examined along the eigen-directions of the local landscape. The three quantities: λ_k, α_k, and μ_k under ℓ_1 normalization (i.e., $\lambda_k/\|\boldsymbol{\lambda}\|_1, \alpha_k/\|\boldsymbol{\alpha}\|_1, \mu_k/\|\boldsymbol{\mu}\|_1$) are plotted. Here, λ_k and α_k represent the curvature and the component of noise energy along the k-th eigen-direction, respectively. μ_k corresponds to the k-th eigenvalue of the noise covariance matrix, which is included for comparison. One can see that the alignment between α_k and λ_k still exists for ResNet-38 and VGG-13, but the ratio between them becomes significantly larger. As a comparison, we refer to Figure 3b, where the ratio is well-controlled for small-scale networks trained for classifying the same dataset. We hypothesize that thiis observation is consistent with our theoretical results in Section 4: one-sided bounds require much less samples.



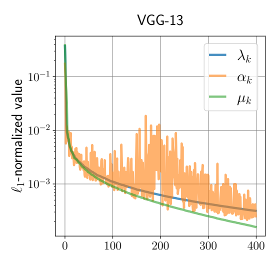


Figure 5: Three distributions ($\{\lambda_k\}_k$, $\{\alpha_k\}_k$, and $\{\mu_k\}_k$) for larger-scale neural networks, which reflect the directional alignment along the eigen directions of the local landscape.

The escape direction of SGD. For large models, it is computationally prohibitive to compute the quantity $D_k(t)$ since it needs to compute the whole spectrum. Thus, we consider to measure the component along different directions without reweighting. Let $\boldsymbol{\theta}^*$ be the minimum of interest and $\boldsymbol{\theta}(t)$ be SGD/GD solution at step t. Define $p_k(t) = \langle \boldsymbol{\theta}(t) - \boldsymbol{\theta}^*, \boldsymbol{u}_1 \rangle$ for $k = 1$ and $p_k(t) = (\sum_{i=1}^k \langle \boldsymbol{\theta}(t) - \boldsymbol{\theta}^*, \boldsymbol{u}_i \rangle^2)^{1/2}$ for $k > 1$; $r_k(t) = (\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}^*\|^2 - p_k^2(t))^{1/2}$. Notably, $p_k(t)$ and $r_k(t)$ represent the component along sharp and flat directions, respectively.

In Figure 6, we plot $(p_k(t), r_k(t))$ for VGG-19 and ResNet-110, where we examine various k values. The plots clearly demonstrate that the escape direction of SGD exhibits significant components along the flat directions. On the other hand, GD tends to escape along much sharper directions. These empirical findings align well with our theoretical findings in Section 5.

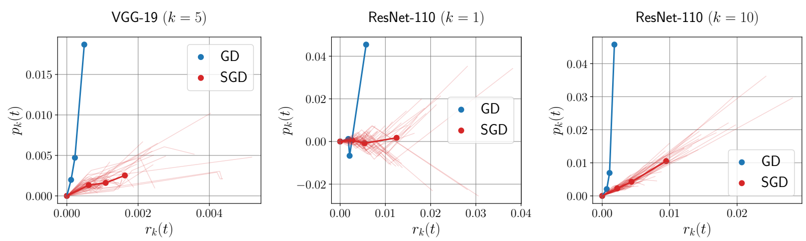


Figure 6: The red curves are 50 escaping trajectories of SGD and their average; the blue curves corresponding to GD. The sharp minimum $\boldsymbol{\theta}^*$ is found by SGD. Then, we run SGD and GD starting from $\boldsymbol{\theta}^*$ and the learning rates are tuned to ensure escaping.

E Conclusion and Future Work

In this paper, we present a comprehensive investigation of the geometry of SGD noise, demonstrating both average and directional alignment between the noise and local geometry. We substantiate these claims through both theoretical analyses and empirical evidence. Furthermore, we explore the implications of these findings by analyzing the escape direction of SGD and its role in enhancing the implicit bias toward flatter minima through cyclical learning rate.

Understanding the noise geometry is crucial for comprehending many aspects of stochastic optimization, including but not limited to convergence rates, generalization capabilities, and dynamic behavior. We offer an illustrative example through analyzing the escape direction of SGD. Another particularly relevant application of our noise geometry framework lies in deciphering the Edge of Stability (EoS) and the associated unstable convergence phenomena, as elaborated below.

- Studies (Cohen et al., 2020; Wu et al., 2018) showed that in training neural networks, GD typically occurs in a EoS phase, where the the stability condition is violated. During EoS phase, GD repeatedly slides into sharp regions and then, escapes from there. Due to the fact that GD escapes along the sharpest direction (as stated in our Proposition 5.3), GD in the EoS phase will keep *oscillating along the sharpest directions* and decreasing the loss along other flat directions. Thus, EoS facilitates the unstable convergence of GD (Ahn et al., 2022). Similar EoS-related phenomena and unstable convergence patterns are also observed in SGD (Lee and Jang, 2022). However, to fully characterize the EoS phase in the context of SGD, it is imperative to understand the underlying noise structure. Specifically, one must elucidate the mechanism by which noise compels SGD to move away from sharp minima.
- In addition, our finding can potentially be used to explain why the training curve of SGD can be more stable than that of GD—A very counter-intuitive phenomenon. As shown in Fig. 2 of Geiping et al. (2021), GD training often encounters *sudden large loss spikes* and in contrast, SGD training does not have this issue (although there are small loss fluctuations), implying that minibatch noise can stabilizes the training to some extent. This can potentially be explained by our theory as follows. For both SGD and GD, the unstable dynamics is inevitable in training neural networks due to progressive sharpening, i.e., entering the EoS phase. During the EoS phase, GD escapes along the sharpest direction, leading to a sudden large loss spike if the curvature along the sharpest direction becomes extremely large. In contrast, for SGD, the escape happens along much flatter directions, for which it is unlikely to trigger a large loss spike.

F Experimental Setups

In this section, we provide the experiment details for directional alignment experiments (in Figure 3 and Figure 5) and escaping experiments (in Figure 4 and Figure 6).

Small-scale experiments (Figure 3 and 4).

- In Figure 3, we conduct experiments on linear regression and a 4-layer linear network: $d \to m \to m \to m \to 1$ with $m = 50$. The inputs $\{\boldsymbol{x}_i\}_{i=1}^n$ are drawn from $\mathcal{N}(\mathbf{0}, I_d)$. In the first three experiments, we fix $d = 10^3$ and change n accordingly $(n = 4d^2, n = d, n = d/4)$.

For the last experiment, we set $d = 10^4$ and $n = \log d$. Regarding the parameter $\boldsymbol{\theta}$, it is drawn from $\mathcal{N}(\mathbf{0}, I_p)$.

- In Figure 4, we conduct escaping experiments on linear regression with $\boldsymbol{w}^* = \mathbf{0}$. Both SGD and GD are initialized near the global minimum by $\boldsymbol{w}(0) \sim \mathcal{N}(\mathbf{0}, e^{-10} I_d/d)$. To ensure escaping, we choose $\eta = 1.2/\|G\|_{\mathrm{F}}$ and $\eta = 4/(\lambda_1 + \lambda_2)$ for SGD and GD, respectively. We fix $n = 10^5$ and $d = 10^3$, and the inputs $\{\boldsymbol{x}_i\}_{i=1}^n$ are drawn from $\mathcal{N}(\mathbf{0}, \mathrm{diag}(\boldsymbol{\lambda})/d)$, where $\boldsymbol{\lambda} \in \mathbb{R}^d$ and $\lambda_1 \geq \lambda_2 = \cdots = \lambda_d \geq 0$. Moreover, we set $\lambda_1 = 1$ change λ_2 accordingly to obtain different $\mathrm{srk}(G^2)$.

Larger-scale experiments (Figure 5 and 6).

- Dataset. For the experiments in Figure 5 and 6, we use the CIFAR-10 dataset with label=$0, 1$ and the full CIFAR-10 dataset to train our models, respectively.
- Models. We conduct experiments on large-scale models: 4-layer CNN ($p = 43,072$), 4-layer FNN ($p = 219,200$), ResNet-38 ($p = 558,222$), VGG-13 ($p = 605,458$), ResNet-110 ($p = 1,720,138$), and VGG-19 ($p = 20,091,338$).

  Specifically, we use standard ResNets (He et al., 2016) and VGG nets (Simonyan and Zisserman, 2015) without batch normalization. For ResNets, we follow Zhang et al. (2019) to use the fixup initialization in order to ensure that the model can be trained without batch normalization. Moreover, the architecture of 4-layer CNN is $\texttt{Conv}(3,6,5) \to \texttt{ReLU} \to \texttt{MPool}(2,2) \to \texttt{Conv}(6,16,5) \to \texttt{ReLU} \to \texttt{MPool}(2,2) \to \texttt{Linear}(400,100) \to \texttt{ReLU} \to \texttt{Linear}(100,2)$. and the 4-layer FNN is a ReLU-activated fully-connected network with the architecture: $784 \to 256 \to 64 \to 32 \to 2$.
- Training. All explicit regularizations (including weight decay, dropout, data augmentation, batch normalization, learning rate decay) are removed, and a simple constant-LR SGD is used to train our models. Specifically, all these models are trained by SGD with learning rate $\eta = 0.1$ and batch size $B = 32$ until the training loss becomes smaller than 10^{-4}.

Efficient computations of the top-k eigen-decomposition of G and Σ. We utilize the functions `eigsh` and `LinearOperator` in `scipy.sparse.linalg` to calculate top-k eigenvalues and eigenvectors of G and Σ, and the key step is to efficiently calculate $G\boldsymbol{v}$ and $\Sigma\boldsymbol{v}$ for any given $\boldsymbol{v} \in \mathbb{R}^p$.

- For small-scale experiments, they can be calculated directly.
- For the large-scale models, we need further approximations since the computation complexity $\mathcal{O}(np)$ is prohibitive in this case. To illustrate our method, we will use $G\boldsymbol{v}$ as an example and apply a similar approach to $\Sigma\boldsymbol{v}$. Notice that the formulation $G\boldsymbol{v} = \frac{1}{n}\sum_{i=1}^n (\boldsymbol{x}_i^\top \boldsymbol{v})\boldsymbol{x}_i$ are all in the form of sample average, which allows us to perform Monte-Carlo approximation. Specifically, we randomly choose b samples $\{\boldsymbol{x}_{i_j}\}_{j=1}^b$ from $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ and use $\frac{1}{b}\sum_{j=1}^b (\boldsymbol{x}_{i_j}^\top \boldsymbol{v})\boldsymbol{x}_{i_j}$ estimate $G\boldsymbol{v}$, with the computation complexity $\mathcal{O}(bp)$. For the experiments on CIFAR-10, we test b's with different values and find that $b = 2k$ is sufficient to obtain a reliable approximation of the top-k eigenvalues and eigenvectors. Hence, for all large-scale experiments in this paper, we use $b = 2k$ to speed up the computation of the top-k eigenvalues and eigenvectors.

G Proofs in Section 3: Average alignment

G.1 Proof of Theorem 3.1 (a)

For clarity, in a slightly different order from the main text, we first prove for the linear model (Example) and then for the OLM (Theorem 3.1). This is also convenient for us to compare the difference between the proof for the two-layer neural network (Theorem 3.4) and the proof for the linear model.

Step I. *Proof for linear models.*

For the linear model, i.e., $\boldsymbol{\theta}=\boldsymbol{w}$ and $\boldsymbol{F}(\boldsymbol{w})=\boldsymbol{w}$ in OLMs, we have

$$\begin{aligned}
\mu(\boldsymbol{w}) &= \frac{\mathrm{Tr}\left(\Sigma(\boldsymbol{w})G(\boldsymbol{w})\right)}{2\mathcal{L}(\boldsymbol{w})\left\|G(\boldsymbol{w})\right\|_{\mathrm{F}}^2} \\
&= \frac{\mathrm{Tr}\left(\left(\frac{1}{n}\sum_{j=1}^n \boldsymbol{x}_j\boldsymbol{x}_j^\top\right)\left(\frac{1}{n}\sum_{i=1}^n (F(\boldsymbol{\theta})^\top\boldsymbol{x}_i)^2(\nabla F(\boldsymbol{\theta})^\top\boldsymbol{x}_i)(\nabla F(\boldsymbol{\theta})^\top\boldsymbol{x}_i)^\top\right)\right)}{\left(\frac{1}{n}\sum_{i=1}^n (F(\boldsymbol{\theta})^\top\boldsymbol{x}_i)^2\right)\left(\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n\left(\boldsymbol{x}_i^\top\nabla F(\boldsymbol{\theta})\nabla F(\boldsymbol{\theta})^\top\boldsymbol{x}_j\right)^2\right)} \\
&= \frac{\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n(\boldsymbol{w}^\top\boldsymbol{x}_i)^2\left(\boldsymbol{x}_i^\top\boldsymbol{x}_j\right)^2}{\left(\frac{1}{n}\sum_{i=1}^n(\boldsymbol{w}^\top\boldsymbol{x}_i)^2\right)\left(\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n\left(\boldsymbol{x}_i^\top\boldsymbol{x}_j\right)^2\right)} \geq \frac{\left(\frac{1}{n}\sum_{i=1}^n(\boldsymbol{w}^\top\boldsymbol{x}_i)^2\right)\left(\min_{i\in[n]}\frac{1}{n}\sum_{j=1}^n\left(\boldsymbol{x}_i^\top\boldsymbol{x}_j\right)^2\right)}{\left(\frac{1}{n}\sum_{i=1}^n(\boldsymbol{w}^\top\boldsymbol{x}_i)^2\right)\left(\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n\left(\boldsymbol{x}_i^\top\boldsymbol{x}_j\right)^2\right)} \qquad (1) \\
&= \frac{\min_{i\in[n]}\frac{1}{n}\sum_{j=1}^n\left(\boldsymbol{x}_i^\top\boldsymbol{x}_j\right)^2}{\max_{i\in[n]}\frac{1}{n}\sum_{j=1}^n\left(\boldsymbol{x}_i^\top\boldsymbol{x}_j\right)^2} \geq \frac{\min_{i\in[n]}\left\|\boldsymbol{x}_i\right\|^4+(n-1)\min_{i\in[n]}\frac{1}{n-1}\sum_{j\neq i}(\boldsymbol{x}_i^\top\boldsymbol{x}_j)^2}{\max_{i\in[n]}\left\|\boldsymbol{x}_i\right\|^4+(n-1)\max_{i\in[n]}\frac{1}{n-1}\sum_{j\neq i}(\boldsymbol{x}_i^\top\boldsymbol{x}_j)^2}.
\end{aligned}$$

Then we only need to estimate $\left\|\boldsymbol{x}_i\right\|^4$ and $\frac{1}{n-1}\sum_{j\neq i}(\boldsymbol{x}_i^\top\boldsymbol{x}_j)^2$ for each $i\in[n]$, respectively.

<u>Step I (i). Estimation of $\left\|\boldsymbol{x}_i\right\|^4$.</u>

Let $\boldsymbol{y}_i=S^{1/2}\boldsymbol{x}_i$, then $\left\|\boldsymbol{x}_i\right\|^2=\boldsymbol{y}_i^\top S\boldsymbol{y}_i$ and $\boldsymbol{y}_1,\cdots,\boldsymbol{y}_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0},I_d)$.

For a fix $i\in[n]$, by Lemma J.2, there exists an absolute constant $C_1>0$ such that for any $\epsilon\in(0,1)$, we have

$$\mathbb{P}\left(\left|\boldsymbol{y}_i^\top S\boldsymbol{y}_i-\mathrm{Tr}(S)\right|\geq\epsilon\mathrm{Tr}(S)\right)\leq 2\exp\left(-C_1\min\left\{\frac{\epsilon^2\mathrm{Tr}^2(S)}{\left\|S\right\|_{\mathrm{F}}^2},\frac{\epsilon\mathrm{Tr}(S)}{\left\|S\right\|_2}\right\}\right).$$

Noticing that $\mathrm{Tr}(S)\left\|S\right\|_2=\lambda_1\sum_i\lambda_i\geq\sum_i\lambda_i^2=\left\|S\right\|_F$, we thus have

$$\frac{\mathrm{Tr}^2(S)}{\left\|S\right\|_{\mathrm{F}}^2}\geq\frac{\mathrm{Tr}(S)}{\left\|S\right\|_2}=\mathrm{srk}(S).$$

Therefore,

$$\mathbb{P}\left(\left|\boldsymbol{y}_i^\top S\boldsymbol{y}_i-\mathrm{Tr}(S)\right|\geq\epsilon\mathrm{Tr}(S)\right)\leq 2\exp\left(-C_1\frac{\mathrm{Tr}(S)}{\left\|S\right\|_2}\min\left\{\epsilon,\epsilon^2\right\}\right)=2\exp\left(-C_1\epsilon^2\mathrm{srk}(S)\right).$$

Applying a union bound over all $i\in[n]$, we have

$$\mathbb{P}\left(\left|\left\|\boldsymbol{x}_i\right\|^2-\mathrm{Tr}(S)\right|\geq\epsilon\mathrm{Tr}(S),\forall i\in[n]\right)\leq 2n\exp\left(-C_1\epsilon^2\mathrm{srk}(S)\right).$$

In the other word, for any $\epsilon,\delta\in(0,1)$, if $\mathrm{srk}(S)\gtrsim\log(n)/\epsilon^2$, then *w.p.* at least $1-\delta/3$, we have

$$(1-\epsilon)^2\leq\frac{\left\|\boldsymbol{x}_i\right\|_2^4}{\mathrm{Tr}^2(S)}\leq(1+\epsilon)^2,\ \forall i\in[n].$$

<u>Step I (ii). Estimation of $\frac{1}{n-1}\sum_{j\neq i}(\boldsymbol{x}_i^\top\boldsymbol{x}_j)^2$.</u>

First, we fix $i\in[n]$. Notice that $(\boldsymbol{x}_i^\top\boldsymbol{x}_j)^2$ $(j\neq i)$ are not independent, so we need estimate by some decoupling tricks.

We denote $\boldsymbol{y}_i:=S^{-1/2}\boldsymbol{x}_i$, then $\boldsymbol{y}_1,\cdots,\boldsymbol{y}_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0},I_d)$ and $(\boldsymbol{x}_i^\top\boldsymbol{x}_j)^2=(\boldsymbol{y}_i^\top S\boldsymbol{y}_j)^2$.

For any fixed $\boldsymbol{v} \in \mathbb{S}^{d-1}$, by Lemma J.1, for any $\epsilon \in (0,1)$, we have

$$\begin{aligned}
&\mathbb{P}\left(\Big|\frac{1}{n-1}\sum_{j\neq i}(\boldsymbol{v}^\top \boldsymbol{y}_j)^2 - 1\Big| \geq \epsilon\right) \\
\leq&\mathbb{P}\left(\Big|\frac{1}{n-1}\sum_{j\neq i}(\boldsymbol{v}^\top \boldsymbol{y}_j)^2 - 1\Big| \geq \epsilon\right) \leq 2\exp\left(-C_2(n-1)\epsilon^2\right),
\end{aligned}$$

where $C_2 > 0$ is an absolute constant, independent of $\boldsymbol{v}$ and ϵ.

Then we have

$$\begin{aligned}
&\mathbb{P}\left(\Big|\frac{1}{n-1}\sum_{j\neq i}(\boldsymbol{x}_i^\top \boldsymbol{x}_j)^2 - \boldsymbol{x}_i^\top S\boldsymbol{x}_i\Big| \geq \epsilon\boldsymbol{x}_i^\top S\boldsymbol{x}_i\right) \\
=&\mathbb{P}\left(\Big|\frac{1}{n-1}\sum_{j\neq i}(\boldsymbol{y}_i^\top S\boldsymbol{y}_j)^2 - \left\|S\boldsymbol{y}_i\right\|_2^2\Big| \geq \epsilon\left\|S\boldsymbol{y}_i\right\|_2^2\right) \\
\overset{\boldsymbol{z}_i:=S\boldsymbol{y}_i/\|S\boldsymbol{y}_i\|_2}{=}&\mathbb{P}\left(\Big|\frac{1}{n-1}\sum_{j\neq i}(\boldsymbol{z}_i^\top \boldsymbol{y}_j)^2 - 1\Big| \geq \epsilon\right) \\
=&\mathbb{E}\left[\mathbb{I}\left\{\Big|\frac{1}{n-1}\sum_{j\neq i}(\boldsymbol{z}_i^\top \boldsymbol{y}_j)^2 - 1\Big| \geq 1\right\}\right] \\
=&\mathbb{E}_{\boldsymbol{z}_i}\left[\mathbb{E}\left[\mathbb{I}\left\{\Big|\frac{1}{n-1}\sum_{j\neq i}(\boldsymbol{z}_i^\top \boldsymbol{y}_j)^2 - 1\Big| \geq 1\right\}\Big|\boldsymbol{z}_i\right]\right] \\
\leq&\mathbb{E}_{\boldsymbol{z}_i}\left[2\exp\left(-C_2(n-1)\epsilon^2\right)\right] = 2\exp\left(-C_2(n-1)\epsilon^2\right).
\end{aligned}$$

Applying a union bound over all $i \in [n]$, we have

$$\mathbb{P}\left(\Big|\frac{1}{n-1}\sum_{j\neq i}(\boldsymbol{x}_i^\top \boldsymbol{x}_j)^2 - \boldsymbol{x}_i^\top S\boldsymbol{x}_i\Big| \geq \epsilon\boldsymbol{x}_i^\top S\boldsymbol{x}_i, \forall i \in [n]\right) \leq 2n\exp\left(-C_2(n-1)\epsilon^2\right).$$

In the other word, for any $\epsilon, \delta \in (0,1)$, if $n/\log(n/\delta) \gtrsim 1/\epsilon^2$, then *w.p.* at least $1-\delta/3$, we have

$$1-\epsilon \leq \frac{\frac{1}{n-1}\sum_{j\neq i}(\boldsymbol{x}_i^\top \boldsymbol{x}_j)^2}{\boldsymbol{x}_i^\top S\boldsymbol{x}_i} \leq 1+\epsilon,\ \forall i \in [n].$$

<u>Step I (iii). Estimation of $\boldsymbol{x}_i^\top S\boldsymbol{x}_i$.</u>

Let $\boldsymbol{y}_i = S^{1/2}\boldsymbol{x}_i$, then $\boldsymbol{x}_i^\top S\boldsymbol{x}_i = \boldsymbol{y}_i^\top S^2\boldsymbol{y}_i$ and $\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, I_d)$.

In the same way as Step I(i), we obtain that: for any $\epsilon, \delta \in (0,1)$, if $\mathrm{srk}(S^2) \gtrsim \log(n)/\epsilon^2$, then *w.p.* at least $1-\delta/3$, we have

$$1-\epsilon \leq \frac{\boldsymbol{x}_i^\top S\boldsymbol{x}_i}{\mathrm{Tr}(S^2)} \leq 1+\epsilon,\ \forall i \in [n].$$

Combining our results in Step I (i), Step I (ii), and Step I (iii), we obtain the result for Linear Model: for any $\epsilon, \delta \in (0,1)$, if $n/\log(n/\delta) \gtrsim 1/\epsilon^2$ and $\min\{\mathrm{srk}(S), \mathrm{srk}(S^2)\} \gtrsim \log(n)/\epsilon^2$, then *w.p.* at least $1-\delta/3-\delta/3-\delta/3 = 1-\delta$, we have

$$\mu(\boldsymbol{w}) \geq \frac{(1-\epsilon)^2\mathrm{Tr}^2(S) + (n-1)(1-\epsilon)\min\limits_{i\in[n]}\boldsymbol{x}_i^\top S\boldsymbol{x}_i}{(1+\epsilon)^2\mathrm{Tr}^2(S) + (n-1)(1+\epsilon)\max\limits_{i\in[n]}\boldsymbol{x}_i^\top S\boldsymbol{x}_i}$$

$$\geq \frac{(1-\epsilon)^2 \mathrm{Tr}^2(S) + (n-1)(1-\epsilon)^2 \mathrm{Tr}(S^2)}{(1+\epsilon)^2 \mathrm{Tr}^2(S) + (n-1)(1+\epsilon)^2 \mathrm{Tr}(S^2)} = \frac{(1-\epsilon)^2}{(1+\epsilon)^2}.$$

From the arbitrary of $\boldsymbol{w}$, we have $\inf_{\boldsymbol{w} \in \mathbb{R}^d} \mu(\boldsymbol{w}) \geq \frac{(1-\epsilon)^2}{(1+\epsilon)^2}$.

Step II. *Proof for OLMs.*

$$\begin{aligned}
&\mu(\boldsymbol{\theta}) = \frac{\mathrm{Tr}\left(\Sigma(\boldsymbol{\theta}) G(\boldsymbol{\theta})\right)}{2\mathcal{L}(\boldsymbol{\theta}) \left\|G(\boldsymbol{\theta})\right\|_{\mathrm{F}}^2} \\
&= \frac{\mathrm{Tr}\left(\left(\frac{1}{n} \sum_{j=1}^{n} (\nabla F(\boldsymbol{\theta})^\top \boldsymbol{x}_j)(\nabla F(\boldsymbol{\theta})^\top \boldsymbol{x}_j)^\top \right) \left(\frac{1}{n} \sum_{i=1}^{n} (F(\boldsymbol{\theta})^\top \boldsymbol{x}_i)^2 (\nabla F(\boldsymbol{\theta})^\top \boldsymbol{x}_i)(\nabla F(\boldsymbol{\theta})^\top \boldsymbol{x}_i)^\top \right) \right)}{\left(\frac{1}{n} \sum_{i=1}^{n} (F(\boldsymbol{\theta})^\top \boldsymbol{x}_i)^2 \right) \left(\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\boldsymbol{x}_i^\top \nabla F(\boldsymbol{\theta}) \nabla F(\boldsymbol{\theta})^\top \boldsymbol{x}_j \right)^2 \right)} \\
&= \frac{\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} (F(\boldsymbol{\theta})^\top \boldsymbol{x}_i)^2 \left(\boldsymbol{x}_i^\top \nabla F(\boldsymbol{\theta}) \nabla F(\boldsymbol{\theta})^\top \boldsymbol{x}_j \right)^2}{\left(\frac{1}{n} \sum_{i=1}^{n} (F(\boldsymbol{\theta})^\top \boldsymbol{x}_i)^2 \right) \left(\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\boldsymbol{x}_i^\top \nabla F(\boldsymbol{\theta}) \nabla F(\boldsymbol{\theta})^\top \boldsymbol{x}_j \right)^2 \right)} \\
&\geq \frac{\left(\frac{1}{n} \sum_{i=1}^{n} (F(\boldsymbol{\theta})^\top \boldsymbol{x}_i)^2 \right) \left(\min_{i \in [n]} \frac{1}{n} \sum_{j=1}^{n} \left(\boldsymbol{x}_i^\top \nabla F(\boldsymbol{\theta}) \nabla F(\boldsymbol{\theta})^\top \boldsymbol{x}_j \right)^2 \right)}{\left(\frac{1}{n} \sum_{i=1}^{n} (F(\boldsymbol{\theta})^\top \boldsymbol{x}_i)^2 \right) \left(\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\boldsymbol{x}_i^\top \nabla F(\boldsymbol{\theta}) \nabla F(\boldsymbol{\theta})^\top \boldsymbol{x}_j \right)^2 \right)} = \frac{\min_{i \in [n]} \frac{1}{n} \sum_{j=1}^{n} \left(\boldsymbol{x}_i^\top \nabla F(\boldsymbol{\theta}) \nabla F(\boldsymbol{\theta})^\top \boldsymbol{x}_j \right)^2}{\max_{i \in [n]} \frac{1}{n} \sum_{j=1}^{n} \left(\boldsymbol{x}_i^\top \nabla F(\boldsymbol{\theta}) \nabla F(\boldsymbol{\theta})^\top \boldsymbol{x}_j \right)^2} \\
&\geq \frac{\min_{i \in [n]} \left\| \nabla F(\boldsymbol{\theta})^\top \boldsymbol{x}_i \right\|^4 + (n-1) \min_{i \in [n]} \frac{1}{n-1} \sum_{j \neq i} (\boldsymbol{x}_i^\top \nabla F(\boldsymbol{\theta}) \nabla F(\boldsymbol{\theta})^\top \boldsymbol{x}_j)^2}{\max_{i \in [n]} \left\| \nabla F(\boldsymbol{\theta})^\top \boldsymbol{x}_i \right\|^4 + (n-1) \max_{i \in [n]} \frac{1}{n-1} \sum_{j \neq i} (\boldsymbol{x}_i^\top \nabla F(\boldsymbol{\theta}) \nabla F(\boldsymbol{\theta})^\top \boldsymbol{x}_j)^2}.
\end{aligned} \tag{2}$$

We can still prove the theorem by the similar way as Step I.

By replacing $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ ($j \neq i$) in Step I (i) with $\nabla F(\boldsymbol{\theta}) \nabla F(\boldsymbol{\theta})^\top \boldsymbol{x}_i$ and $\boldsymbol{x}_j$ ($j \neq i$), respectively, in the similar way as Step I (i), we can obtain: for any $\epsilon, \delta \in (0,1)$, if $n / \log(n/\delta) \gtrsim 1/\epsilon^2$, then *w.p.* at least $1-\delta$, we have

$$1 - \epsilon \leq \frac{\frac{1}{n-1} \sum_{j \neq i} (\boldsymbol{x}_i^\top \nabla F(\boldsymbol{\theta}) \nabla F(\boldsymbol{\theta})^\top \boldsymbol{x}_j)^2}{\boldsymbol{x}_i^\top \nabla F(\boldsymbol{\theta}) \nabla F(\boldsymbol{\theta})^\top S \nabla F(\boldsymbol{\theta}) \nabla F(\boldsymbol{\theta})^\top \boldsymbol{x}_i} \leq 1 + \epsilon, \ \forall i \in [n];$$

Combining the estimation above with Step I (ii) and Step I (iii), we obtain that: for any $\epsilon, \delta \in (0,1)$, if $n / \log(n/\delta) \gtrsim 1/\epsilon^2$ and $\mathrm{srk}(S^2) \gtrsim \log(n)/\epsilon^2$, then *w.p.* at least $1-\delta$, we have

$$1 - \epsilon \leq \frac{\frac{1}{n-1} \sum_{j \neq i} (\boldsymbol{x}_i^\top \nabla F(\boldsymbol{\theta}) \nabla F(\boldsymbol{\theta})^\top \boldsymbol{x}_j)^2}{\boldsymbol{x}_i^\top \nabla F(\boldsymbol{\theta}) \nabla F(\boldsymbol{\theta})^\top S \nabla F(\boldsymbol{\theta}) \nabla F(\boldsymbol{\theta})^\top \boldsymbol{x}_i} \leq 1 + \epsilon, \ \forall i \in [n];$$

$$(1-\epsilon)^2 \leq \frac{\|\boldsymbol{x}_i\|_2^4}{\mathrm{Tr}^2(S)} \leq (1+\epsilon)^2, \ \forall i \in [n];$$

$$1 - \epsilon \leq \frac{\boldsymbol{x}_i^\top S \boldsymbol{x}_i}{\mathrm{Tr}(S^2)} \leq 1 + \epsilon, \ \forall i \in [n].$$

These inequalities imply that:

$$\begin{aligned}
&\mu(\boldsymbol{\theta}) \geq \frac{\min_{i \in [n]} \lambda_{\min}^2(\nabla F(\boldsymbol{\theta}) \nabla F(\boldsymbol{\theta})^\top) \|\boldsymbol{x}_i\|_2^4 + (n-1) \min_{i \in [n]} \frac{1}{n-1} \sum_{j \neq i} (\boldsymbol{x}_i^\top \nabla F(\boldsymbol{\theta}) \nabla F(\boldsymbol{\theta})^\top \boldsymbol{x}_j)^2}{\max_{i \in [n]} \lambda_{\min}^2(\nabla F(\boldsymbol{\theta}) \nabla F(\boldsymbol{\theta})^\top) \|\boldsymbol{x}_i\|_2^4 + (n-1) \max_{i \in [n]} \frac{1}{n-1} \sum_{j \neq i} (\boldsymbol{x}_i^\top \nabla F(\boldsymbol{\theta}) \nabla F(\boldsymbol{\theta})^\top \boldsymbol{x}_j)^2} \\
&\geq \frac{(1-\epsilon)^2 \min_{i \in [n]} \lambda_{\min}^2(\nabla F(\boldsymbol{\theta}) \nabla F(\boldsymbol{\theta})^\top) \mathrm{Tr}^2(S) + (n-1)(1-\epsilon) \min_{i \in [n]} \boldsymbol{x}_i^\top \nabla F(\boldsymbol{\theta}) \nabla F(\boldsymbol{\theta})^\top S \nabla F(\boldsymbol{\theta}) \nabla F(\boldsymbol{\theta})^\top \boldsymbol{x}_i}{(1-\epsilon)^2 \max_{i \in [n]} \lambda_{\max}^2(\nabla F(\boldsymbol{\theta}) \nabla F(\boldsymbol{\theta})^\top) \mathrm{Tr}^2(S) + (n-1)(1+\epsilon) \max_{i \in [n]} \boldsymbol{x}_i^\top \nabla F(\boldsymbol{\theta}) \nabla F(\boldsymbol{\theta})^\top S \nabla F(\boldsymbol{\theta}) \nabla F(\boldsymbol{\theta})^\top \boldsymbol{x}_i}
\end{aligned}$$

$$
\geq \frac{(1-\epsilon)^2 \min_{i\in[n]} \lambda_{\min}^2(\nabla F(\boldsymbol{\theta})\nabla F(\boldsymbol{\theta})^\top)\mathrm{Tr}^2(S) + (n-1)(1-\epsilon)\lambda_{\min}^2(\nabla F(\boldsymbol{\theta})\nabla F(\boldsymbol{\theta})^\top) \min_{i\in[n]} \boldsymbol{x}_i^\top S \boldsymbol{x}_i}{(1+\epsilon)^2 \max_{i\in[n]} \lambda_{\max}^2(\nabla F(\boldsymbol{\theta})\nabla F(\boldsymbol{\theta})^\top)\mathrm{Tr}^2(S) + (n-1)(1+\epsilon)\lambda_{\max}^2(\nabla F(\boldsymbol{\theta})\nabla F(\boldsymbol{\theta})^\top) \max_{i\in[n]} \boldsymbol{x}_i^\top S \boldsymbol{x}_i}
$$

$$
\geq \frac{(1-\epsilon)^2 \min_{i\in[n]} \lambda_{\min}^2(\nabla F(\boldsymbol{\theta})\nabla F(\boldsymbol{\theta})^\top)\mathrm{Tr}^2(S) + (n-1)(1-\epsilon)^2\lambda_{\min}^2(\nabla F(\boldsymbol{\theta})\nabla F(\boldsymbol{\theta})^\top)\mathrm{Tr}(S^2)}{(1+\epsilon)^2 \max_{i\in[n]} \lambda_{\max}^2(\nabla F(\boldsymbol{\theta})\nabla F(\boldsymbol{\theta})^\top)\mathrm{Tr}^2(S) + (n-1)(1+\epsilon)^2\lambda_{\max}^2(\nabla F(\boldsymbol{\theta})\nabla F(\boldsymbol{\theta})^\top)\mathrm{Tr}(S^2)}
$$

$$
= \frac{(1-\epsilon)^2}{(1+\epsilon)^2 \mathrm{cond}^2(\nabla F(\boldsymbol{\theta})\nabla F(\boldsymbol{\theta})^\top)}.
$$

Hence, we have proved Theorem 3.1. □

G.2 Proof of Theorem 3.1 (b)

This result is a direct corollary of Theorem 4.2, which is proved in Appendix H.

Under the same setting as Theorem 4.2, Theorem 4.2 gives us the uniform lower bound: there exists an absolute constant $C > 0$ such that

$$
\inf_{\boldsymbol{\theta},\boldsymbol{v}\in\mathbb{R}^p} g(\boldsymbol{\theta};\boldsymbol{v}) \geq C,
$$

which means that for any $\boldsymbol{\theta} \in \mathbb{R}^p, \boldsymbol{v} \in \mathbb{S}^{p-1}$, we have

$$
\boldsymbol{v}^\top \Sigma(\boldsymbol{\theta})\boldsymbol{v} \geq C \cdot 2\mathcal{L}(\boldsymbol{\theta})\boldsymbol{v}^\top G(\boldsymbol{\theta})\boldsymbol{v}.
$$

Consider the orthogonal decomposition of $G(\boldsymbol{\theta})$: $G(\boldsymbol{\theta}) = \sum_{k=1}^p \lambda_k \boldsymbol{u}_k \boldsymbol{u}_k^\top$. Notice that

$$
\mathrm{Tr}(\Sigma(\boldsymbol{\theta})G(\boldsymbol{\theta})) = \sum_{k=1}^p \lambda_k \boldsymbol{u}_k^\top \Sigma(\boldsymbol{\theta})\boldsymbol{u}_k,
$$

$$
\|G(\boldsymbol{\theta})\|_{\mathrm{F}} = \mathrm{Tr}(G(\boldsymbol{\theta})G(\boldsymbol{\theta})) = \sum_{k=1}^p \lambda_k \boldsymbol{u}_k^\top G(\boldsymbol{\theta})\boldsymbol{u}_k.
$$

Then we obtain

$$
\mathrm{Tr}(\Sigma(\boldsymbol{\theta})G(\boldsymbol{\theta})) \geq C \cdot 2\mathcal{L}(\boldsymbol{\theta}) \sum_{k=1}^p \lambda_k \boldsymbol{u}_k^\top G(\boldsymbol{\theta})\boldsymbol{u}_k = C \cdot 2\mathcal{L}(\boldsymbol{\theta}) \|G(\boldsymbol{\theta})\|_{\mathrm{F}}^2,
$$

which means $\mu(\boldsymbol{\theta}) \geq C$. From the arbitrariness of $\boldsymbol{\theta}$, it holds that $\inf_{\boldsymbol{\theta}\in\mathbb{R}^p} \mu(\boldsymbol{\theta}) \geq C$. □

G.3 Proof of Theorem 3.4

For two-layer neural networks with fixed output layer, the gradient is

$$
\nabla f(\boldsymbol{x}_i;\boldsymbol{\theta}) = \left(a_1\sigma'(\boldsymbol{b}_1^\top \boldsymbol{x}_i)\boldsymbol{x}_i^\top, \cdots, a_m\sigma'(\boldsymbol{b}_m^\top \boldsymbol{x}_i)\boldsymbol{x}_i^\top\right)^\top \in \mathbb{R}^{md}.
$$

For simplicity, denote $\nabla f_i(\boldsymbol{\theta}) := \nabla f(\boldsymbol{x}_i;\boldsymbol{\theta})$, $\boldsymbol{u}_i(\boldsymbol{\theta}) := f_i(\boldsymbol{\theta}) - f_i(\boldsymbol{\theta}^*)$. Then we have:

$$
\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2n}\sum_{i=1}^n u_i^2(\boldsymbol{\theta}), \quad G(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n \nabla f_i(\boldsymbol{\theta})\nabla f_i(\boldsymbol{\theta})^\top, \quad \Sigma(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n u_i^2(\boldsymbol{\theta})\nabla f_i(\boldsymbol{\theta})\nabla f_i(\boldsymbol{\theta})^\top.
$$

$$
\mu(\boldsymbol{\theta}) = \frac{\mathrm{Tr}\left(\left(\frac{1}{n}\sum_{i=1}^n \nabla f_i(\boldsymbol{\theta})\nabla f_i(\boldsymbol{\theta})^\top\right)\left(\frac{1}{n}\sum_{i=1}^n u_i^2(\boldsymbol{\theta})\nabla f_i(\boldsymbol{\theta})\nabla f_i(\boldsymbol{\theta})^\top\right)\right)}{\left(\frac{1}{n}\sum_{i=1}^n u_i^2(\boldsymbol{\theta})\right)\left(\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \left(\nabla f_i(\boldsymbol{\theta})^\top \nabla f_i(\boldsymbol{\theta})\right)^2\right)}
$$

$$
\begin{aligned}
&= \frac{\frac{1}{n}\sum_{i=1}^n u_i^2(\boldsymbol{\theta})\frac{1}{n}\sum_{j=1}^n \left(\nabla f_i(\boldsymbol{\theta})^\top \nabla f_j(\boldsymbol{\theta})\right)^2}{\left(\frac{1}{n}\sum_{i=1}^n u_i^2(\boldsymbol{\theta})\right)\left(\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n (\nabla f_i(\boldsymbol{\theta})^\top \nabla f_i(\boldsymbol{\theta}))^2\right)} \\
&\geq \frac{\min_{i\in[n]}\frac{1}{n}\sum_{j=1}^n \left(\nabla f_i(\boldsymbol{\theta})^\top \nabla f_j(\boldsymbol{\theta})\right)^2}{\frac{1}{n^2}\sum_{i=1}^n\sum_{i=1}^n (\nabla f_i(\boldsymbol{\theta})^\top \nabla f_j(\boldsymbol{\theta}))^2} \geq \frac{\min_{i\in[n]}\frac{1}{n}\sum_{j=1}^n \left(\alpha^2 m \boldsymbol{x}_i^\top \boldsymbol{x}_j\right)^2}{\frac{1}{n^2}\sum_{i=1}^n\sum_{i=1}^n \left(\beta^2 m \boldsymbol{x}_i^\top \boldsymbol{x}_j\right)^2} = \frac{\alpha^2}{\beta^2}\frac{\min_{i\in[n]}\frac{1}{n}\sum_{j=1}^n \left(\boldsymbol{x}_i^\top \boldsymbol{x}_j\right)^2}{\frac{1}{n^2}\sum_{i=1}^n\sum_{i=1}^n \left(\boldsymbol{x}_i^\top \boldsymbol{x}_j\right)^2}.
\end{aligned}
$$

565 Notice that the last term $\frac{\min_{i\in[n]}\frac{1}{n}\sum_{j=1}^n \left(\boldsymbol{x}_i^\top \boldsymbol{x}_j\right)^2}{\frac{1}{n^2}\sum_{i=1}^n\sum_{i=1}^n \left(\boldsymbol{x}_i^\top \boldsymbol{x}_j\right)^2}$ is independent of $\boldsymbol{\theta}$ and the same as (1) for the linear model . Then repeating the same proof of Linear Model, the result of this theorem differs from Linear Model by only the factor α^2/β^2. In other words, under the same condition with Linear Model, *w.p.* at least $1-\delta$, we have

$$
\inf_{\boldsymbol{\theta}\in\mathbb{R}^{md}} \mu(\boldsymbol{\theta}) \geq \frac{\alpha^2}{\beta^2}\frac{(1-\epsilon)^2}{(1+\epsilon)^2}.
$$

□

H Proofs in Section 4: Directional Alignment

For the OLM $f(\boldsymbol{x};\boldsymbol{\theta}) = F(\boldsymbol{\theta})^T\boldsymbol{x}$, let $\boldsymbol{r}(\boldsymbol{\theta}) = F(\boldsymbol{\theta}) - F(\boldsymbol{\theta}^*)$. Then, we have

$$
\begin{aligned}
\hat{G}(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i=1}^n \nabla F^\top(\boldsymbol{\theta})\boldsymbol{x}_i\boldsymbol{x}_i^\top \nabla F(\boldsymbol{\theta}) \\
\hat{\mathcal{L}}(\boldsymbol{\theta}) &= \frac{1}{2n}\sum_{i=1}^n \left(\boldsymbol{u}^\top(\boldsymbol{\theta})\boldsymbol{x}_i\right)^2 \qquad (3)\\
\hat{\Sigma}(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i=1}^n \left(\boldsymbol{r}^\top(\boldsymbol{\theta})\boldsymbol{x}_i\right)^2 \nabla F^\top(\boldsymbol{\theta})\boldsymbol{x}_i\boldsymbol{x}_i^\top \nabla F(\boldsymbol{\theta}),
\end{aligned}
$$

and for the population case:

$$
\begin{aligned}
G(\boldsymbol{\theta}) &= \mathbb{E}\left[\nabla F^\top(\boldsymbol{\theta})\boldsymbol{x}\boldsymbol{x}^\top \nabla F(\boldsymbol{\theta})\right] = \nabla F^\top(\boldsymbol{\theta}) S \nabla F(\boldsymbol{\theta}) \\
\mathcal{L}(\boldsymbol{\theta}) &= \frac{1}{2}\mathbb{E}\left[\left(\boldsymbol{r}^\top(\boldsymbol{\theta})\boldsymbol{x}\right)^2\right] = \frac{1}{2}\boldsymbol{r}(\boldsymbol{\theta})^\top S \boldsymbol{r}(\boldsymbol{\theta}) \\
\Sigma(\boldsymbol{\theta}) &= \mathbb{E}\left[\left(\boldsymbol{r}^\top(\boldsymbol{\theta})\boldsymbol{x}\right)^2 \nabla F^\top(\boldsymbol{\theta})\boldsymbol{x}\boldsymbol{x}^\top \nabla F(\boldsymbol{\theta})\right]
\end{aligned}
$$

Lemma H.1 (Proposition 2.3 in (Wu et al., 2022))**.** *Let the data distribution be $\mathcal{N}(\mathbf{0}, S)$. Then we have*

$$
\Sigma(\boldsymbol{\theta}) = \nabla\mathcal{L}(\boldsymbol{\theta})\nabla\mathcal{L}(\boldsymbol{\theta})^\top + 2\mathcal{L}(\boldsymbol{\theta})G(\boldsymbol{\theta}).
$$

573 **Lemma H.2.** *Under the same conditions in Lemma H.1, if $\boldsymbol{u}(\boldsymbol{\theta}) \neq \mathbf{0}$ and $\nabla F(\boldsymbol{\theta})\boldsymbol{v} \neq 0$, then we have:*

$$
\left(\nabla\mathcal{L}(\boldsymbol{\theta})^\top \boldsymbol{v}\right)^2 \leq 2\mathcal{L}(\boldsymbol{\theta})\boldsymbol{v}^\top G(\boldsymbol{\theta})\boldsymbol{v}.
$$

Proof. Noticing that $\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\boldsymbol{r}(\boldsymbol{\theta})^\top S \boldsymbol{r}(\boldsymbol{\theta})$, we have $\nabla\mathcal{L}(\boldsymbol{\theta}) = \nabla F(\boldsymbol{\theta})^\top S \boldsymbol{u}(\boldsymbol{\theta})$. Hence,

$$
\begin{aligned}
\left(\nabla\mathcal{L}(\boldsymbol{\theta})^\top \boldsymbol{v}\right)^2 &= \boldsymbol{v}^\top \nabla F(\boldsymbol{\theta})^\top S \boldsymbol{r}(\boldsymbol{\theta})\boldsymbol{r}(\boldsymbol{\theta})^\top S \nabla F(\boldsymbol{\theta})\boldsymbol{v} = \langle \nabla F(\boldsymbol{\theta})\boldsymbol{v}, \boldsymbol{r}(\boldsymbol{\theta})\rangle_S^2 \\
&\overset{\text{Lemma J.6}}{\leq} \|\nabla F(\boldsymbol{\theta})\boldsymbol{v}\|_S^2 \|\boldsymbol{r}(\boldsymbol{\theta})\|_S^2 = 2\mathcal{L}(\boldsymbol{\theta})\left(\boldsymbol{v}\nabla F(\boldsymbol{\theta})^\top S \nabla F(\boldsymbol{\theta})\boldsymbol{v}\right) = 2\mathcal{L}(\boldsymbol{\theta})\boldsymbol{v}^\top G(\boldsymbol{\theta})\boldsymbol{v}.
\end{aligned}
$$

□

Lemma H.3. *Let* $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$. *For any* $\epsilon, \delta \in (0,1)$, *if we choose* $n \gtrsim (d + \log(1/\delta))/\epsilon^2$, *then* w.p. *at least* $1-\delta$, *we have:*

$$\sup_{\boldsymbol{v} \in \mathbb{S}^{d-1}} \left| \frac{1}{n} \sum_{i=1}^n (\boldsymbol{v}^\top \boldsymbol{x}_i)^2 - 1 \right| \leq \epsilon.$$

Proof. By Lemma J.3 with $K = \sqrt{C_1}$, we know that: *w.p.* at least $1 - 2\exp(-u)$, we have

$$\left\| \frac{1}{n} \sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^\top - \boldsymbol{I}_d \right\| \leq C_2 \left(\sqrt{\frac{d+u}{n}} + \frac{d+u}{n} \right),$$

where C_2 is an absolute positive constant. Equivalently, we can rewrite this conclusion. For any $\epsilon, \delta \in (0,1)$, if we choose $n \gtrsim (d + \log(1/\delta))/\epsilon^2$, then *w.p.* at least $1-\delta$, we have:

$$\sup_{\boldsymbol{v} \in \mathbb{S}^{d-1}} \left| \frac{1}{n} \sum_{i=1}^n (\boldsymbol{v}^\top \boldsymbol{x}_i)^2 - 1 \right| \leq \left\| \frac{1}{n} \sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^\top - \boldsymbol{I}_d \right\| \leq \epsilon.$$

□

Lemma H.4 (Corollary 2 in (Cai et al., 2022))**.** *Let* $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$. *There exists absolute constants* $C_1, C_2, C_3 > 0$, *such that if* $n \geq C_3 d$, *then* w.p. *at least* $1 - \exp(-C_2 n)$, *we have*

$$\inf_{\boldsymbol{u}, \boldsymbol{v} \in \mathbb{S}^{d-1}} \frac{1}{n} \sum_{i=1}^n (\boldsymbol{x}_i^\top \boldsymbol{u})^2 (\boldsymbol{x}_i^\top \boldsymbol{v})^2 \geq C_1.$$

With the preparation of Lemma H.3 and Lemma H.4, now we give the proof of Theorem 4.2.

H.1 Proof of Theorem 4.2

Let $\boldsymbol{y}_i = \boldsymbol{S}^{-1/2} \boldsymbol{x}_i$, then $\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, I_d)$.

$$\begin{aligned} g(\boldsymbol{\theta}; \boldsymbol{v}) = &= \frac{\frac{1}{n} \sum\limits_{i=1}^n \left(\boldsymbol{r}^\top(\boldsymbol{\theta}) \boldsymbol{x}_i \right)^2 \left(\left(\nabla F(\boldsymbol{\theta}) \boldsymbol{v} \right)^\top \boldsymbol{x}_i \right)^2}{\frac{1}{n} \sum\limits_{i=1}^n \left(\boldsymbol{r}^\top(\boldsymbol{\theta}) \boldsymbol{x}_i \right)^2 \cdot \frac{1}{n} \sum\limits_{i=1}^n \left(\left(\nabla F(\boldsymbol{\theta}) \boldsymbol{v} \right)^\top \boldsymbol{x}_i \right)^2} \\ &= \frac{\frac{1}{n} \sum\limits_{i=1}^n \left((\boldsymbol{S}^{1/2} \boldsymbol{r}(\boldsymbol{\theta}))^\top \boldsymbol{y}_i \right)^2 \left(\left(\boldsymbol{S}^{1/2} \nabla F(\boldsymbol{\theta}) \boldsymbol{v} \right)^\top \boldsymbol{y}_i \right)^2}{\frac{1}{n} \sum\limits_{i=1}^n \left((\boldsymbol{S}^{1/2} \boldsymbol{r}(\boldsymbol{\theta}))^\top \boldsymbol{y}_i \right)^2 \cdot \frac{1}{n} \sum\limits_{i=1}^n \left(\left(\boldsymbol{S}^{1/2} \nabla F(\boldsymbol{\theta}) \boldsymbol{v} \right)^\top \boldsymbol{y}_i \right)^2}, \end{aligned}$$

Case(i). If $\boldsymbol{S}^{1/2} \boldsymbol{r}(\boldsymbol{\theta}) = \mathbf{0}$ or $\boldsymbol{S}^{1/2} \nabla F(\boldsymbol{\theta}) \boldsymbol{v} = 0$, we have $g(\boldsymbol{\theta}; \boldsymbol{v}) = \frac{0}{0} = 1$, this theorem holds.

Case (ii). If $\boldsymbol{S}^{1/2} \boldsymbol{r}(\boldsymbol{\theta}) \neq \mathbf{0}$ and $\boldsymbol{S}^{1/2} \nabla F(\boldsymbol{\theta}) \boldsymbol{v} \neq 0$, we define the following normalized vectors:

$$\tilde{\boldsymbol{r}}(\boldsymbol{\theta}) := \frac{\boldsymbol{S}^{1/2} \boldsymbol{r}(\boldsymbol{\theta})}{\left\| \boldsymbol{S}^{1/2} \boldsymbol{r}(\boldsymbol{\theta}) \right\|} \in \mathbb{S}^{d-1} \quad \tilde{\boldsymbol{w}}(\boldsymbol{\theta}; \boldsymbol{v}) := \frac{\boldsymbol{S}^{1/2} \nabla F(\boldsymbol{\theta}) \boldsymbol{v}}{\left\| \boldsymbol{S}^{1/2} \nabla F(\boldsymbol{\theta}) \boldsymbol{v} \right\|} \in \mathbb{S}^{d-1}.$$

From the homogeneity of $g(\boldsymbol{\theta}; \boldsymbol{v})$, we have:

$$g(\boldsymbol{\theta}; \boldsymbol{v}) = \frac{\frac{1}{n} \sum\limits_{i=1}^n \left(\tilde{\boldsymbol{r}}(\boldsymbol{\theta})^\top \boldsymbol{y}_i \right)^2 \left(\tilde{\boldsymbol{w}}(\boldsymbol{\theta}; \boldsymbol{v})^\top \boldsymbol{y}_i \right)^2}{\frac{1}{n} \sum\limits_{i=1}^n \left(\tilde{\boldsymbol{r}}(\boldsymbol{\theta})^\top \boldsymbol{y}_i \right)^2 \cdot \frac{1}{n} \sum\limits_{i=1}^n \left(\tilde{\boldsymbol{w}}(\boldsymbol{\theta}; \boldsymbol{v})^\top \boldsymbol{y}_i \right)^2}.$$

One the one hand, with the help of Lemma H.4, there exists $C_1 > 0$ such that if we choose $n \gtrsim d + \log(1/\delta)$, then *w.p.* at least $1 - \delta/2$, we have:

$$\inf_{\boldsymbol{w}, \boldsymbol{u} \in \mathbb{S}^{d-1}} \frac{1}{n} \sum_{i=1}^n (\boldsymbol{w}^\top \boldsymbol{y}_i)^2 (\boldsymbol{u}^\top \boldsymbol{y}_i)^2 \geq C_1.$$

On the other hand, with the help of Lemma H.3, if we choose $\epsilon = 1/2$ and $n \gtrsim d + \log(1/\delta)$, then *w.p.* at least $1 - \delta/2$, we have:

$$\sup_{\boldsymbol{w}\in\mathbb{S}^{d-1}} \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}^\top \boldsymbol{y}_i)^2 \geq 1 + \frac{1}{2} = \frac{3}{2},$$

Combining these two bounds, we obtain that: if we choose $\epsilon = 1/2$ and $n \gtrsim d + \log(1/\delta)$, then *w.p.* at least $1 - \delta$, we have:

$$\inf_{\boldsymbol{w},\boldsymbol{u}\in\mathbb{S}^{d-1}} \frac{\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}^\top \boldsymbol{y}_i)^2(\boldsymbol{u}^\top \boldsymbol{y}_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}^\top \boldsymbol{y}_i)^2 \cdot \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{u}^\top \boldsymbol{y}_i)^2}$$
$$\geq \frac{\inf_{\boldsymbol{w},\boldsymbol{u}\in\mathbb{S}^{d-1}} \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}^\top \boldsymbol{y}_i)^2(\boldsymbol{u}^\top \boldsymbol{y}_i)^2}{\left(\sup_{\boldsymbol{w}\in\mathbb{S}^{d-1}} \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}^\top \boldsymbol{y}_i)^2\right)^2} \geq \frac{4C_1}{9},$$

which implies that

$$\inf_{\boldsymbol{\theta},\boldsymbol{v}\in\mathbb{R}^p} g(\boldsymbol{\theta};\boldsymbol{v}) \geq \min\left\{1, \inf_{\boldsymbol{w},\boldsymbol{u}\in\mathbb{S}^{d-1}} \frac{\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}^\top \boldsymbol{y}_i)^2(\boldsymbol{u}^\top \boldsymbol{y}_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}^\top \boldsymbol{y}_i)^2 \cdot \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{u}^\top \boldsymbol{y}_i)^2}\right\} \geq \min\left\{1, \frac{4C_1}{9}\right\}.$$

□

H.2 Proof of Theorem 4.3

We first need a few lemmas.

Lemma H.5. *Let* $\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$. *If* $n \gtrsim d^2 + \log^2(1/\delta)$*, then* w.p. *at least* $1 - \delta$*, we have*

$$\sup_{\boldsymbol{v}\in\mathbb{S}^{d-1}} \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{y}_i^\top \boldsymbol{v})^4 \leq 8.$$

Proof. For $\mathbb{S}^{d-1}$, its covering number has the bound:

$$\left(\frac{1}{\rho}\right)^d \leq \mathcal{N}(\mathbb{S}^{d-1}, \rho) \leq \left(\frac{2}{\rho} + 1\right)^d,$$

so there exist a ρ-net on $\mathbb{S}^{d-1}$: $\mathcal{V} \subset \mathbb{S}^{d-1}$, s.t. $|\mathcal{V}| \leq \left(\frac{2}{\rho} + 1\right)^d$.

Step I. Bounding the term on the ρ-net.

For a fixed $\boldsymbol{v} \in \mathcal{V}$, due to $\boldsymbol{y}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$, we can verify $(\boldsymbol{y}_i^\top \boldsymbol{v})^4$ is sub-Weibull random variable:

$$\mathbb{E}\exp\left(\left((\boldsymbol{y}_i^\top \boldsymbol{v})^4\right)^{1/2}\right) = \mathbb{E}\exp\left((\boldsymbol{y}_i^\top \boldsymbol{v})^2\right) \lesssim 1,$$

which means that there exist an absolute constant $C_1 \geq 1$ s.t. $\left\|(\boldsymbol{y}_i^\top \boldsymbol{v})^4\right\|_{\psi_{1/2}} \leq C_1$.

By the concentration inequality for Sub-Weibull distribution with $\beta = 1/2$ (Lemma J.5) and $\mathbb{E}\left[(\boldsymbol{y}^\top \boldsymbol{v})^4\right] = 3$, there exists an absolute constant $C_2 \geq 1$ s.t.

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\left[(\boldsymbol{y}_i^\top \boldsymbol{v})^4\right] - 3\right| > \phi(n;\delta)\right) \leq 2\delta,$$

where $\phi(n;\delta) = C_2(\sqrt{\frac{\log(1/\delta)}{n}} + \frac{\log^2(1/\delta)}{n})$. Applying a union bound over $\boldsymbol{v} \in \mathcal{V}$, we have:

$$\mathbb{P}\left(\exists \boldsymbol{v} \in \mathcal{V} s.t. \left|\frac{1}{n}\sum_{i=1}^{n}\left[(\boldsymbol{y}_i^\top \boldsymbol{v})^4\right] - 3\right| > \phi(n;\delta)\right)$$

$$\leq \mathbb{P}\left(\bigcup_{\boldsymbol{v}\in\mathcal{V}}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\left[(\boldsymbol{y}_i^\top \boldsymbol{v})^4\right]-3\right|>\phi(n;\delta)\right\}\right)\leq \sum_{\boldsymbol{v}\in\mathcal{V}}\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\left[(\boldsymbol{y}_i^\top \boldsymbol{v})^4\right]-3\right|>\phi(n;\delta)\right)$$
$$\leq 2|\mathcal{V}|\exp\left(-\frac{n}{C_2^2}\right)=2\left(\frac{2}{\rho}+1\right)^d\delta.$$

So *w.p.* at least $1-2\left(\frac{2}{\rho}+1\right)^d\delta$, we have:

$$\max_{\boldsymbol{v}\in\mathcal{V}}\frac{1}{n}\sum_{i=1}^{n}\left[(\boldsymbol{y}_i^\top \boldsymbol{v})^4\right]\leq 3+\phi(n;\delta).$$

Step II. Estimate the error of the ρ-net approximation.

For simplicity, we denote

$$P:=\max_{\boldsymbol{v}\in\mathbb{S}^{d-1}}\frac{1}{n}\sum_{i=1}^{n}\left[(\boldsymbol{y}_i^\top \boldsymbol{v})^4\right]\quad ,Q:=\max_{\boldsymbol{v}\in\mathcal{V}}\frac{1}{n}\sum_{i=1}^{n}\left[(\boldsymbol{y}_i^\top \boldsymbol{v})^4\right].$$

Let $\boldsymbol{v}\in\mathbb{S}^{d-1}$ such that $\frac{1}{n}\sum_{i=1}^n\left[(\boldsymbol{y}_i^\top \boldsymbol{v})^4\right]=P$, then there exist $\boldsymbol{v}_0\in\mathcal{V}$, s.t. $\|\boldsymbol{v}-\boldsymbol{v}_0\|\leq\rho$.

On the one hand,

$$\left|\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{y}_i^\top \boldsymbol{v})^4-\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{y}_i^\top \boldsymbol{v}_0)^4\right|=\left|\frac{1}{n}\sum_{i=1}^{n}\left((\boldsymbol{y}_i^\top \boldsymbol{v})^4-(\boldsymbol{y}_i^\top \boldsymbol{v}_0)^4\right)\right|$$
$$=\left|\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{y}_i^\top(\boldsymbol{v}-\boldsymbol{v}_0)\right)\left(\boldsymbol{y}_i^\top(\boldsymbol{v}+\boldsymbol{v}_0)\right)\left((\boldsymbol{y}_i^\top \boldsymbol{v})^2+(\boldsymbol{y}_i^\top \boldsymbol{v}_0)^2\right)\right|$$
$$\leq\left|\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{y}_i^\top(\boldsymbol{v}-\boldsymbol{v}_0)\right)\left(\boldsymbol{y}_i^\top(\boldsymbol{v}+\boldsymbol{v}_0)\right)(\boldsymbol{y}_i^\top \boldsymbol{v})^2\right|+\left|\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{y}_i^\top(\boldsymbol{v}-\boldsymbol{v}_0)\right)\left(\boldsymbol{y}_i^\top(\boldsymbol{v}+\boldsymbol{v}_0)\right)(\boldsymbol{y}_i^\top \boldsymbol{v}_0)^2\right|$$
$$\leq\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{y}_i^\top(\boldsymbol{v}-\boldsymbol{v}_0)\right)^2\left(\boldsymbol{y}_i^\top(\boldsymbol{v}+\boldsymbol{v}_0)\right)^2}\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{y}_i^\top \boldsymbol{v})^4}+\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{y}_i^\top \boldsymbol{v}_0)^4}\right)$$
$$\leq\sqrt[4]{\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{y}_i^\top(\boldsymbol{v}-\boldsymbol{v}_0)\right)^4}\sqrt[4]{\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{y}_i^\top(\boldsymbol{v}+\boldsymbol{v}_0)\right)^4}\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{y}_i^\top \boldsymbol{v})^4}+\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{y}_i^\top \boldsymbol{v}_0)^4}\right)$$
$$\leq\|\boldsymbol{v}-\boldsymbol{v}_0\|P^{1/4}\|\boldsymbol{v}+\boldsymbol{v}_0\|P^{1/4}(\sqrt{P}+\sqrt{Q})\leq 2\rho\sqrt{P}(\sqrt{P}+\sqrt{Q})$$

On the other hand,

$$\left|\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{y}_i^\top \boldsymbol{v})^4-\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{y}_i^\top \boldsymbol{v}_0)^4\right|\geq P-\sum_{i=1}^{n}(\boldsymbol{y}_i^\top \boldsymbol{v}_0)^4\geq P-Q.$$

Hence, we obtain

$$P-Q\leq 2\rho\sqrt{P}(\sqrt{P}+\sqrt{Q}),$$

which means that

$$P\leq\left(\frac{1}{1-2\rho}\right)^2Q.$$

Step III. The bound for any $\boldsymbol{v}\in\mathbb{S}^{d-1}$.

Select $\rho=\frac{1}{2}(1-\frac{1}{\sqrt{2}})$ and denote $\delta'=2(\frac{2}{\rho}+1)^d\delta$. And we choose $n\gtrsim d^2+\log^2(1/\delta')$, which ensures $\phi(n;\delta)\leq 1$.

Then combining the results in Step I and Step II, we know that: *w.p.* at least $1-\delta'$, we have:

$$\max_{\boldsymbol{v}\in\mathcal{V}}\frac{1}{n}\sum_{i=1}^{n}\left[(\boldsymbol{y}_i^\top\boldsymbol{v})^4\right]\leq 3+1=4;\quad \max_{\boldsymbol{v}\in\mathbb{S}^{d-1}}\frac{1}{n}\sum_{i=1}^{n}\left[(\boldsymbol{y}_i^\top\boldsymbol{v})^4\right]\leq 2\max_{\boldsymbol{v}\in\mathcal{V}}\frac{1}{n}\sum_{i=1}^{n}\left[(\boldsymbol{y}_i^\top\boldsymbol{v})^4\right],$$

which means

$$\max_{\boldsymbol{v}\in\mathbb{S}^{d-1}}\frac{1}{n}\sum_{i=1}^{n}\left[(\boldsymbol{y}_i^\top\boldsymbol{v})^4\right]\leq 2\cdot 4=8.$$

□

Lemma H.6. *Let* $\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n\overset{\text{i.i.d.}}{\sim}\mathcal{N}(\mathbf{0},\boldsymbol{I}_d)$. *For any* $\epsilon,\delta\in(0,1)$, *if we choose*

$$n\gtrsim\max\left\{\left(d^2\log^2\left(1/\epsilon\right)+\log^2(1/\delta)\right)/\epsilon,\left(d\log\left(1/\epsilon\right)+\log(1/\delta)\right)/\epsilon^2\right\},$$

then w.p. *at least* $1-\delta$, *we have:*

$$\sup_{\boldsymbol{w},\boldsymbol{v}\in\mathbb{S}^{d-1}}\left|\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}^\top\boldsymbol{x}_i)^2(\boldsymbol{v}^\top\boldsymbol{x}_i)^2-\mathbb{E}\Big[(\boldsymbol{w}^\top\boldsymbol{x}_1)^2(\boldsymbol{v}^\top\boldsymbol{x}_1)^2\Big]\right|\leq\epsilon.$$

Proof. For $\mathbb{S}^{d-1}$, its covering number has the bound:

$$\left(\frac{1}{\rho}\right)^d\leq\mathcal{N}(\mathbb{S}^{d-1},\rho)\leq\left(\frac{2}{\rho}+1\right)^d,$$

so there exist two ρ-nets on $\mathbb{S}^{d-1}$: $\mathcal{W}\subset\mathbb{S}^{d-1}$ and $\mathcal{V}\subset\mathbb{S}^{d-1}$, s.t.

$$|\mathcal{W}|\leq\left(\frac{2}{\rho}+1\right)^d,\quad |\mathcal{V}|\leq\left(\frac{2}{\rho}+1\right)^d.$$

Step I. Bounding the term on the ρ-net.

In this step, will estimate the term $\left|\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}^\top\boldsymbol{x}_i)^2(\boldsymbol{v}^\top\boldsymbol{x}_i)^2-\mathbb{E}\Big[(\boldsymbol{w}^\top\boldsymbol{x})^2(\boldsymbol{v}^\top\boldsymbol{x})^2\Big]\right|$ for any $\boldsymbol{w}\in\mathcal{W}$ and $\boldsymbol{v}\in\mathcal{V}$.

For fixed $\boldsymbol{w}\in\mathcal{W}$ and $\boldsymbol{v}\in\mathcal{V}$, we denote $X_i^{\boldsymbol{w},\boldsymbol{v}}:=(\boldsymbol{w}^\top\boldsymbol{x}_i)^2(\boldsymbol{v}^\top\boldsymbol{x}_i)^2$. We can verify X_i is a sub-Weibull random variable with $\beta=1/2$ (Definition J.4):

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\left|(\boldsymbol{w}^\top\boldsymbol{x}_i)^2(\boldsymbol{v}^\top\boldsymbol{x}_i)\right|^{1/2}\right)\right]=\mathbb{E}\left[\exp\left(|\boldsymbol{w}^\top\boldsymbol{x}_i||\boldsymbol{v}^\top\boldsymbol{x}_i|\right)\right]\\
\leq&\mathbb{E}\left[\exp\left(\frac{(\boldsymbol{w}^\top\boldsymbol{x}_i)^2+(\boldsymbol{v}^\top\boldsymbol{x}_i)^2}{2}\right)\right]=\mathbb{E}\left[\exp\left(\frac{(\boldsymbol{w}^\top\boldsymbol{x}_i)^2}{2}\right)\exp\left(\frac{(\boldsymbol{v}^\top\boldsymbol{x}_i)^2}{2}\right)\right]\\
\overset{\text{Lemma J.6}}{\leq}&\sqrt{\mathbb{E}\left[\exp\left((\boldsymbol{w}^\top\boldsymbol{x}_i)^2\right)\cdot\sqrt{\mathbb{E}\left[\exp\left((\boldsymbol{v}^\top\boldsymbol{x}_i)^2\right)\right]}\right.}\overset{\|(\boldsymbol{v}^\top\boldsymbol{x}_i)^2\|_{\psi_1}\leq C_3}{\lesssim}1,
\end{aligned}$$

which means that there exists an absolute constant $C_4\geq 1$, s.t. $\|X_i^{\boldsymbol{w},\boldsymbol{v}}\|_{\psi_{1/2}}\leq C_4$. By the concentration inequality for Sub-Weibull distribution with $\beta=1/2$ (Lemma J.5), there exists an absolute constant $C_5\geq 1$, s.t.

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}X_i^{\boldsymbol{w},\boldsymbol{v}}-\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\big[X_i^{\boldsymbol{w},\boldsymbol{v}}\big]\right|>\psi(n;\delta)\right)\leq\delta.$$

where $\psi(n;\delta)=C_5\left(\sqrt{\frac{\log(1/\delta)}{n}}+\frac{(\log(1/\delta))^2}{n}\right)$.

Applying an union bound over $\boldsymbol{w} \in \mathcal{W}$ and $\boldsymbol{v} \in \mathcal{V}$, we have:

$$
\begin{aligned}
&\mathbb{P}\left(\exists \boldsymbol{w} \in \mathcal{W}, \boldsymbol{v} \in \mathcal{V}, \text{s.t.} \left|\frac{1}{n}\sum_{i=1}^{n} X_i^{\boldsymbol{w},\boldsymbol{v}} - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\big[X_i^{\boldsymbol{w},\boldsymbol{v}}\big]\right| > \psi(n;\delta)\right) \\
\leq &\mathbb{P}\left(\bigcup_{(\boldsymbol{w},\boldsymbol{v})\in\mathcal{W}\times\mathcal{V}} \left\{\exists \boldsymbol{w} \in \mathcal{W}, \boldsymbol{v} \in \mathcal{V}, \text{s.t.} \left|\frac{1}{n}\sum_{i=1}^{n} X_i^{\boldsymbol{w},\boldsymbol{v}} - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\big[X_i^{\boldsymbol{w},\boldsymbol{v}}\big]\right| > \psi(n;\delta)\right\}\right) \\
\leq &\sum_{(\boldsymbol{w},\boldsymbol{v})\in\mathcal{W}\times\mathcal{V}} \mathbb{P}\left(\exists \boldsymbol{w} \in \mathcal{W}, \boldsymbol{v} \in \mathcal{V}, \text{s.t.} \left|\frac{1}{n}\sum_{i=1}^{n} X_i^{\boldsymbol{w},\boldsymbol{v}} - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\big[X_i^{\boldsymbol{w},\boldsymbol{v}}\big]\right| > \psi(n;\delta)\right) \\
\leq &2|\mathcal{W}||\mathcal{V}|\delta \leq 2\left(\frac{2}{\rho}+1\right)^{2d}\delta.
\end{aligned}
$$

So *w.p.* at least $1 - 2\left(\frac{2}{\rho}+1\right)^{2d}\delta$, we have:

$$
\sup_{\boldsymbol{w}\in\mathcal{W},\boldsymbol{v}\in\mathcal{V}} \left|\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}^\top\boldsymbol{x}_i)^2(\boldsymbol{v}^\top\boldsymbol{x}_i)^2 - \mathbb{E}\Big[(\boldsymbol{w}^\top\boldsymbol{x})^2(\boldsymbol{v}^\top\boldsymbol{x})^2\Big]\right| \leq \psi(n;\delta).
$$

<u>Step II. Estimate the population error of the ρ-net approximation.</u>

Let $\boldsymbol{w}, \boldsymbol{v}, \boldsymbol{w}_0, \boldsymbol{v}_0 \in \mathbb{S}^{d-1}$, s.t. $\|\boldsymbol{w}-\boldsymbol{w}_0\| \leq \rho$ and $\|\boldsymbol{v}-\boldsymbol{v}_0\| \leq \rho$. For the population error, we have

$$
\begin{aligned}
&\left|\mathbb{E}\Big[(\boldsymbol{w}^\top\boldsymbol{x})^2(\boldsymbol{v}^\top\boldsymbol{x})^2\Big] - \mathbb{E}\Big[(\boldsymbol{w}_0^\top\boldsymbol{x})^2(\boldsymbol{v}_0^\top\boldsymbol{x})^2\Big]\right| \\
=&\left|\mathbb{E}\Big[\Big((\boldsymbol{w}^\top\boldsymbol{x})^2 - (\boldsymbol{w}_0^\top\boldsymbol{x})^2\Big)(\boldsymbol{v}^\top\boldsymbol{x})^2\Big] + \mathbb{E}\Big[(\boldsymbol{w}_0^\top\boldsymbol{x})^2\Big((\boldsymbol{v}^\top\boldsymbol{x})^2 - (\boldsymbol{v}_0^\top\boldsymbol{x})^2\Big)\Big]\right| \\
\leq&\left|\mathbb{E}\Big[\Big((\boldsymbol{w}^\top\boldsymbol{x})^2 - (\boldsymbol{w}_0^\top\boldsymbol{x})^2\Big)(\boldsymbol{v}^\top\boldsymbol{x})^2\Big]\right| + \left|\mathbb{E}\Big[(\boldsymbol{w}_0^\top\boldsymbol{x})^2\Big((\boldsymbol{v}^\top\boldsymbol{x})^2 - (\boldsymbol{v}_0^\top\boldsymbol{x})^2\Big)\Big]\right|
\end{aligned}
$$

We first bound $\left|\mathbb{E}\Big[\Big((\boldsymbol{w}^\top\boldsymbol{x})^2 - (\boldsymbol{w}_0^\top\boldsymbol{x})^2\Big)(\boldsymbol{v}^\top\boldsymbol{x})^2\Big]\right|$:

$$
\begin{aligned}
&\left|\mathbb{E}\Big[\Big((\boldsymbol{w}^\top\boldsymbol{x})^2 - (\boldsymbol{w}_0^\top\boldsymbol{x})^2\Big)(\boldsymbol{v}^\top\boldsymbol{x})^2\Big]\right| = \left|\mathbb{E}\Big[\Big((\boldsymbol{w}-\boldsymbol{w}_0)^\top\boldsymbol{x}\boldsymbol{x}^\top(\boldsymbol{w}+\boldsymbol{w}_0)(\boldsymbol{v}^\top\boldsymbol{x})^2\Big]\right| \\
\leq&\Big(\mathbb{E}\left[\left((\boldsymbol{w}-\boldsymbol{w}_0)^\top\boldsymbol{x}\boldsymbol{x}^\top(\boldsymbol{w}+\boldsymbol{w}_0)\right)^2\right]\Big)^{1/2}\left(\mathbb{E}\left[(\boldsymbol{v}^\top\boldsymbol{x})^4\right]\right)^{1/2} \\
\leq&\Big(\mathbb{E}\left[\left((\boldsymbol{w}-\boldsymbol{w}_0)^\top\boldsymbol{x}\right)^4\right]\Big)^{1/4}\Big(\mathbb{E}\left[\left((\boldsymbol{w}+\boldsymbol{w}_0)^\top\boldsymbol{x}\right)^4\right]\Big)^{1/4}\left(\mathbb{E}\left[(\boldsymbol{v}^\top\boldsymbol{x})^4\right]\right)^{1/2} \\
\leq&3\left\|(\boldsymbol{w}-\boldsymbol{w}_0)\right\|\left\|(\boldsymbol{w}+\boldsymbol{w}_0)\right\|\left\|\boldsymbol{v}\right\|^2 \leq 6\rho.
\end{aligned}
$$

Repeating the proof above, we also have:

$$
\left|\mathbb{E}\Big[\Big((\boldsymbol{w}^\top\boldsymbol{x})^2 - (\boldsymbol{w}_0^\top\boldsymbol{x})^2\Big)(\boldsymbol{v}^\top\boldsymbol{x})^2\Big]\right| \leq 6\rho.
$$

Combining these two inequalities, we have:

$$
\left|\mathbb{E}\Big[(\boldsymbol{w}^\top\boldsymbol{x})^2(\boldsymbol{v}^\top\boldsymbol{x})^2\Big] - \mathbb{E}\Big[(\boldsymbol{w}_0^\top\boldsymbol{x})^2(\boldsymbol{v}_0^\top\boldsymbol{x})^2\Big]\right| \leq 6\rho + 6\rho = 12\rho.
$$

Due to the arbitrariness of $\boldsymbol{w}, \boldsymbol{v}, \boldsymbol{w}_0, \boldsymbol{v}_0$, we obtain

$$
\sup_{\substack{\boldsymbol{w},\boldsymbol{v},\boldsymbol{w}_0,\boldsymbol{v}_0\in\mathbb{S}^{d-1} \\ \|\boldsymbol{w}-\boldsymbol{w}_0\|\leq\rho, \|\boldsymbol{v}-\boldsymbol{v}_0\|\leq\rho}} \left|\mathbb{E}\Big[(\boldsymbol{w}^\top\boldsymbol{x})^2(\boldsymbol{v}^\top\boldsymbol{x})^2\Big] - \mathbb{E}\Big[(\boldsymbol{w}_0^\top\boldsymbol{x})^2(\boldsymbol{v}_0^\top\boldsymbol{x})^2\Big]\right| \leq 12\rho.
$$

<u>Step III. Estimate the empirical error of the ρ-net approximation.</u>

Let $\boldsymbol{w}, \boldsymbol{v}, \boldsymbol{w}_0, \boldsymbol{v}_0 \in \mathbb{S}^{d-1}$, s.t. $\|\boldsymbol{w}-\boldsymbol{w}_0\| \leq \rho$ and $\|\boldsymbol{v}-\boldsymbol{v}_0\| \leq \rho$. For the empirical error, we have

$$
\left|\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}^\top\boldsymbol{x}_i)^2(\boldsymbol{v}^\top\boldsymbol{x}_i)^2 - \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}_0^\top\boldsymbol{x}_i)^2(\boldsymbol{v}_0^\top\boldsymbol{x}_i)^2\right|
$$

$$=\left|\frac{1}{n}\sum_{i=1}^{n}\left[\Big((\boldsymbol{w}^\top\boldsymbol{x}_i)^2-(\boldsymbol{w}_0^\top\boldsymbol{x}_i)^2\Big)(\boldsymbol{v}^\top\boldsymbol{x}_i)^2\right]+\frac{1}{n}\sum_{i=1}^{n}\left[(\boldsymbol{w}_0^\top\boldsymbol{x}_i)^2\Big((\boldsymbol{v}^\top\boldsymbol{x}_i)^2-(\boldsymbol{v}_0^\top\boldsymbol{x}_i)^2\Big)\right]\right|$$
$$\leq\left|\frac{1}{n}\sum_{i=1}^{n}\left[\Big((\boldsymbol{w}^\top\boldsymbol{x}_i)^2-(\boldsymbol{w}_0^\top\boldsymbol{x}_i)^2\Big)(\boldsymbol{v}^\top\boldsymbol{x}_i)^2\right]\right|+\left|\frac{1}{n}\sum_{i=1}^{n}\left[(\boldsymbol{w}_0^\top\boldsymbol{x}_i)^2\Big((\boldsymbol{v}^\top\boldsymbol{x}_i)^2-(\boldsymbol{v}_0^\top\boldsymbol{x}_i)^2\Big)\right]\right|$$

We first bound $\left|\frac{1}{n}\sum_{i=1}^{n}\left[\Big((\boldsymbol{w}^\top\boldsymbol{x}_i)^2-(\boldsymbol{w}_0^\top\boldsymbol{x}_i)^2\Big)(\boldsymbol{v}^\top\boldsymbol{x}_i)^2\right]\right|$:

$$\left|\frac{1}{n}\sum_{i=1}^{n}\left[\Big((\boldsymbol{w}^\top\boldsymbol{x}_i)^2-(\boldsymbol{w}_0^\top\boldsymbol{x}_i)^2\Big)(\boldsymbol{v}^\top\boldsymbol{x}_i)^2\right]\right|=\left|\frac{1}{n}\sum_{i=1}^{n}\left[\Big((\boldsymbol{w}-\boldsymbol{w}_0)^\top\boldsymbol{x}_i\boldsymbol{x}_i^\top(\boldsymbol{w}+\boldsymbol{w}_0)(\boldsymbol{v}^\top\boldsymbol{x}_i)^2\right]\right|$$
$$\leq 2\rho\sup_{\boldsymbol{u}\in\mathbb{S}^{d-1}}\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i^\top\boldsymbol{u})^4.$$

Repeating the proof above, we also have $\left|\frac{1}{n}\sum_{i=1}^{n}\left[(\boldsymbol{w}_0^\top\boldsymbol{x}_i)^2\Big((\boldsymbol{v}^\top\boldsymbol{x}_i)^2-(\boldsymbol{v}_0^\top\boldsymbol{x}_i)^2\Big)\right]\right| \leq 2\rho\sup_{\boldsymbol{u}\in\mathbb{S}^{d-1}}\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i^\top\boldsymbol{u})^4$. Combining these two bounds, we have:

$$\left|\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}^\top\boldsymbol{x}_i)^2(\boldsymbol{v}^\top\boldsymbol{x}_i)^2-\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}_0^\top\boldsymbol{x}_i)^2(\boldsymbol{v}_0^\top\boldsymbol{x}_i)^2\right|\leq 4\rho\sup_{\boldsymbol{u}\in\mathbb{S}^{d-1}}\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i^\top\boldsymbol{u})^4.$$

Using Lemma H.5, if $n \gtrsim d^2+\log^2(1/\delta')$, then *w.p.* at least $1-\delta'/2$, we have $\sup_{\boldsymbol{u}\in\mathbb{S}^{d-1}}\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i^\top\boldsymbol{u})^4\leq 8$.

Hence, *w.p.* at least $1-\delta'/2$, we have

$$\left|\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}^\top\boldsymbol{x}_i)^2(\boldsymbol{v}^\top\boldsymbol{x}_i)^2-\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}_0^\top\boldsymbol{x}_i)^2(\boldsymbol{v}_0^\top\boldsymbol{x}_i)^2\right|\leq 32\rho.$$

Due to the arbitrariness of $\boldsymbol{w},\boldsymbol{v},\boldsymbol{w}_0,\boldsymbol{v}_0$, we obtain

$$\sup_{\substack{\boldsymbol{w},\boldsymbol{v},\boldsymbol{w}_0,\boldsymbol{v}_0\in\mathbb{S}^{d-1}\\ \|\boldsymbol{w}-\boldsymbol{w}_0\|\leq\rho,\|\boldsymbol{v}-\boldsymbol{v}_0\|\leq\rho}}\left|\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}^\top\boldsymbol{x}_i)^2(\boldsymbol{v}^\top\boldsymbol{x}_i)^2-\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}_0^\top\boldsymbol{x}_i)^2(\boldsymbol{v}_0^\top\boldsymbol{x}_i)^2\right|\leq 32\rho.$$

Step IV. The bound for any $\boldsymbol{w},\boldsymbol{v}\in\mathbb{S}^{d-1}$.

Combining the results in Step I, II, and II, we know that *w.p.* at least $1-\frac{\delta'}{2}-(\frac{2}{\rho}+1)^d$, we have

$$\sup_{\boldsymbol{w}\in\mathcal{W},\boldsymbol{v}\in\mathcal{V}}\left|\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}^\top\boldsymbol{x}_i)^2(\boldsymbol{v}^\top\boldsymbol{x}_i)^2-\mathbb{E}\Big[(\boldsymbol{w}^\top\boldsymbol{x})^2(\boldsymbol{v}^\top\boldsymbol{x})^2\Big]\right|\leq\psi(n;\delta),$$
$$\sup_{\substack{\boldsymbol{w},\boldsymbol{v},\boldsymbol{w}_0,\boldsymbol{v}_0\in\mathbb{S}^{d-1}\\ \|\boldsymbol{w}-\boldsymbol{w}_0\|\leq\rho,\|\boldsymbol{v}-\boldsymbol{v}_0\|\leq\rho}}\left|\mathbb{E}\Big[(\boldsymbol{w}^\top\boldsymbol{x})^2(\boldsymbol{v}^\top\boldsymbol{x})^2\Big]-\mathbb{E}\Big[(\boldsymbol{w}_0^\top\boldsymbol{x})^2(\boldsymbol{v}_0^\top\boldsymbol{x})^2\Big]\right|\leq 12\rho,$$
$$\sup_{\substack{\boldsymbol{w},\boldsymbol{v},\boldsymbol{w}_0,\boldsymbol{v}_0\in\mathbb{S}^{d-1}\\ \|\boldsymbol{w}-\boldsymbol{w}_0\|\leq\rho,\|\boldsymbol{v}-\boldsymbol{v}_0\|\leq\rho}}\left|\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}^\top\boldsymbol{x}_i)^2(\boldsymbol{v}^\top\boldsymbol{x}_i)^2-\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}_0^\top\boldsymbol{x}_i)^2(\boldsymbol{v}_0^\top\boldsymbol{x}_i)^2\right|\leq 32\rho.$$

Then for any $\boldsymbol{w},\boldsymbol{v}\in\mathbb{S}^{d-1}$, there exists $\boldsymbol{w}_0\in\mathcal{W},\boldsymbol{v}_0\in\mathcal{V}$ s.t. $\|\boldsymbol{w}-\boldsymbol{w}_0\|\leq\rho$ and $\|\boldsymbol{v}-\boldsymbol{v}_0\|\leq\rho$, so

$$\left|\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}^\top\boldsymbol{x}_i)^2(\boldsymbol{v}^\top\boldsymbol{x}_i)^2-\mathbb{E}\Big[(\boldsymbol{w}^\top\boldsymbol{x})^2(\boldsymbol{v}^\top\boldsymbol{x})^2\Big]\right|$$
$$=\left|\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}^\top\boldsymbol{x}_i)^2(\boldsymbol{v}^\top\boldsymbol{x}_i)^2-\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}_0^\top\boldsymbol{x}_i)^2(\boldsymbol{v}_0^\top\boldsymbol{x}_i)^2+\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}_0^\top\boldsymbol{x}_i)^2(\boldsymbol{v}_0^\top\boldsymbol{x}_i)^2\right.$$

$$
\begin{aligned}
&\quad - \mathbb{E}\Big[(\boldsymbol{w}_0^\top \boldsymbol{x})^2(\boldsymbol{v}_0^\top \boldsymbol{x})^2\Big] + \mathbb{E}\Big[(\boldsymbol{w}_0^\top \boldsymbol{x})^2(\boldsymbol{v}_0^\top \boldsymbol{x})^2\Big] - \mathbb{E}\Big[(\boldsymbol{w}^\top \boldsymbol{x})^2(\boldsymbol{v}^\top \boldsymbol{x})^2\Big]\Bigg| \\
\leq & \left|\frac{1}{n}\sum_{i=1}^n (\boldsymbol{w}^\top \boldsymbol{x}_i)^2(\boldsymbol{v}^\top \boldsymbol{x}_i)^2 - \frac{1}{n}\sum_{i=1}^n (\boldsymbol{w}_0^\top \boldsymbol{x}_i)^2(\boldsymbol{v}_0^\top \boldsymbol{x}_i)^2\right| \\
& + \left|\frac{1}{n}\sum_{i=1}^n (\boldsymbol{w}_0^\top \boldsymbol{x}_i)^2(\boldsymbol{v}_0^\top \boldsymbol{x}_i)^2 - \mathbb{E}\Big[(\boldsymbol{w}_0^\top \boldsymbol{x})^2(\boldsymbol{v}_0^\top \boldsymbol{x})^2\Big]\right| + \left|\mathbb{E}\Big[(\boldsymbol{w}_0^\top \boldsymbol{x})^2(\boldsymbol{v}_0^\top \boldsymbol{x})^2\Big] - \mathbb{E}\Big[(\boldsymbol{w}^\top \boldsymbol{x})^2(\boldsymbol{v}^\top \boldsymbol{x})^2\Big]\right| \\
\leq & \sup_{\substack{\boldsymbol{w},\boldsymbol{v},\boldsymbol{w}_0,\boldsymbol{v}_0\in\mathbb{S}^{d-1} \\ \|\boldsymbol{w}-\boldsymbol{w}_0\|\leq\rho, \|\boldsymbol{v}-\boldsymbol{v}_0\|\leq\rho}} \left|\frac{1}{n}\sum_{i=1}^n (\boldsymbol{w}^\top \boldsymbol{x}_i)^2(\boldsymbol{v}^\top \boldsymbol{x}_i)^2 - \frac{1}{n}\sum_{i=1}^n (\boldsymbol{w}_0^\top \boldsymbol{x}_i)^2(\boldsymbol{v}_0^\top \boldsymbol{x}_i)^2\right| \\
& + \sup_{\boldsymbol{w}\in\mathcal{W},\boldsymbol{v}\in\mathcal{V}} \left|\frac{1}{n}\sum_{i=1}^n (\boldsymbol{w}^\top \boldsymbol{x}_i)^2(\boldsymbol{v}^\top \boldsymbol{x}_i)^2 - \mathbb{E}\Big[(\boldsymbol{w}^\top \boldsymbol{x})^2(\boldsymbol{v}^\top \boldsymbol{x})^2\Big]\right| \\
& + \sup_{\substack{\boldsymbol{w},\boldsymbol{v},\boldsymbol{w}_0,\boldsymbol{v}_0\in\mathbb{S}^{d-1} \\ \|\boldsymbol{w}-\boldsymbol{w}_0\|\leq\rho, \|\boldsymbol{v}-\boldsymbol{v}_0\|\leq\rho}} \left|\mathbb{E}\Big[(\boldsymbol{w}^\top \boldsymbol{x})^2(\boldsymbol{v}^\top \boldsymbol{x})^2\Big] - \mathbb{E}\Big[(\boldsymbol{w}_0^\top \boldsymbol{x})^2(\boldsymbol{v}_0^\top \boldsymbol{x})^2\Big]\right| \\
\leq & 32\rho + \psi(n;\delta) + 12\rho = 44\rho + \psi(n;\delta).
\end{aligned}
$$

Due to the arbitrariness of $\boldsymbol{w}, \boldsymbol{v}$, we have

$$
\sup_{\boldsymbol{w},\boldsymbol{v}\in\mathbb{S}^{d-1}} \left|\frac{1}{n}\sum_{i=1}^n (\boldsymbol{w}^\top \boldsymbol{x}_i)^2(\boldsymbol{v}^\top \boldsymbol{x}_i)^2 - \mathbb{E}\Big[(\boldsymbol{w}^\top \boldsymbol{x})^2(\boldsymbol{v}^\top \boldsymbol{x})^2\Big]\right| \leq 44\rho + \psi(n;\delta)
$$

Select $\rho = \frac{\epsilon}{66}$ and $\delta'/2 = 2(1+\frac{2}{\rho})^{2d}\delta$. And we choose

$$
n \gtrsim \max\left\{\left(d^2\log^2(1/\epsilon) + \log^2(1/\delta)\right)/\epsilon, \left(d\log(1/\epsilon) + \log(1/\delta)\right)/\epsilon^2\right\},
$$

which satisfies $\psi(n;\delta) \leq \epsilon/3$.

Then *w.p.* at least $1 - \delta'/2 - \delta'/2 = 1 - \delta'$, we have

$$
\sup_{\boldsymbol{w},\boldsymbol{v}\in\mathbb{S}^{d-1}} \left|\frac{1}{n}\sum_{i=1}^n (\boldsymbol{w}^\top \boldsymbol{x}_i)^2(\boldsymbol{v}^\top \boldsymbol{x}_i)^2 - \mathbb{E}\Big[(\boldsymbol{w}^\top \boldsymbol{x})^2(\boldsymbol{v}^\top \boldsymbol{x})^2\Big]\right| \leq \frac{44}{66}\epsilon + \frac{1}{3}\epsilon = \epsilon.
$$

□

With the preparation of Lemma H.1, H.3, and H.6, now we give the proof of Theorem 4.3.

Proof of Theorem 4.3. Let $\boldsymbol{y}_i = \boldsymbol{S}^{-1/2}\boldsymbol{x}_i$, then $\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, I_d)$.

$$
\begin{aligned}
g(\boldsymbol{\theta};\boldsymbol{v}) = & \frac{\frac{1}{n}\sum_{i=1}^n \left(\boldsymbol{r}^\top(\boldsymbol{\theta})\boldsymbol{x}_i\right)^2 \left(\left(\nabla F(\boldsymbol{\theta})\boldsymbol{v}\right)^\top \boldsymbol{x}_i\right)^2}{\frac{1}{n}\sum_{i=1}^n \left(\boldsymbol{r}^\top(\boldsymbol{\theta})\boldsymbol{x}_i\right)^2 \cdot \frac{1}{n}\sum_{i=1}^n \left(\left(\nabla F(\boldsymbol{\theta})\boldsymbol{v}\right)^\top \boldsymbol{x}_i\right)^2} \\
= & \frac{\frac{1}{n}\sum_{i=1}^n \left((\boldsymbol{S}^{1/2}\boldsymbol{r}(\boldsymbol{\theta}))^\top \boldsymbol{y}_i\right)^2 \left(\left(\boldsymbol{S}^{1/2}\nabla F(\boldsymbol{\theta})\boldsymbol{v}\right)^\top \boldsymbol{y}_i\right)^2}{\frac{1}{n}\sum_{i=1}^n \left((\boldsymbol{S}^{1/2}\boldsymbol{r}(\boldsymbol{\theta}))^\top \boldsymbol{y}_i\right)^2 \cdot \frac{1}{n}\sum_{i=1}^n \left(\left(\boldsymbol{S}^{1/2}\nabla F(\boldsymbol{\theta})\boldsymbol{v}\right)^\top \boldsymbol{y}_i\right)^2},
\end{aligned}
$$

Case (i). If $\boldsymbol{S}^{1/2}\boldsymbol{r}(\boldsymbol{\theta}) = \mathbf{0}$ or $\boldsymbol{S}^{1/2}\nabla F(\boldsymbol{\theta})\boldsymbol{v} = 0$, we have $g(\boldsymbol{\theta};\boldsymbol{v}) = \frac{0}{0} = 1$, this theorem holds.

Case (ii). If $\boldsymbol{S}^{1/2}\boldsymbol{r}(\boldsymbol{\theta}) \neq \mathbf{0}$ and $\boldsymbol{S}^{1/2}\nabla F(\boldsymbol{\theta})\boldsymbol{v} \neq 0$, we define the following normalized vectors:

$$
\tilde{\boldsymbol{r}}(\boldsymbol{\theta}) := \frac{\boldsymbol{S}^{1/2}\boldsymbol{r}(\boldsymbol{\theta})}{\left\|\boldsymbol{S}^{1/2}\boldsymbol{r}(\boldsymbol{\theta})\right\|} \in \mathbb{S}^{d-1} \quad \tilde{\boldsymbol{w}}(\boldsymbol{\theta};\boldsymbol{v}) := \frac{\boldsymbol{S}^{1/2}\nabla F(\boldsymbol{\theta})\boldsymbol{v}}{\left\|\boldsymbol{S}^{1/2}\nabla F(\boldsymbol{\theta})\boldsymbol{v}\right\|} \in \mathbb{S}^{d-1}.
$$

From the homogeneity of $g(\boldsymbol{\theta};\boldsymbol{v})$, we have:

$$g(\boldsymbol{\theta};\boldsymbol{v})=\frac{\frac{1}{n}\sum_{i=1}^{n}\left(\tilde{\boldsymbol{r}}(\boldsymbol{\theta})^\top\boldsymbol{y}_i\right)^2\left(\tilde{\boldsymbol{w}}(\boldsymbol{\theta};\boldsymbol{v})^\top\boldsymbol{y}_i\right)^2}{\frac{1}{n}\sum_{i=1}^{n}\left(\tilde{\boldsymbol{r}}(\boldsymbol{\theta})^\top\boldsymbol{y}_i\right)^2\cdot\frac{1}{n}\sum_{i=1}^{n}\left(\tilde{\boldsymbol{w}}(\boldsymbol{\theta};\boldsymbol{v})^\top\boldsymbol{y}_i\right)^2}.$$

By Lemma H.3 and H.6, for any $\epsilon,\delta\in(0,1)$, if we choose

$$n\gtrsim\max\left\{\left(d^2\log^2\left(1/\epsilon\right)+\log^2(1/\delta)\right)/\epsilon,\left(d\log\left(1/\epsilon\right)+\log(1/\delta)\right)/\epsilon^2\right\},$$

then *w.p.* at least $1-\delta$, the following inequalities hold:

$$\sup_{\boldsymbol{v}\in\mathbb{S}^{d-1}}\left|\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{v}^\top\boldsymbol{y}_i)^2-1\right|\leq\epsilon,$$
$$\sup_{\boldsymbol{w},\boldsymbol{v}\in\mathbb{S}^{d-1}}\left|\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}^\top\boldsymbol{y}_i)^2(\boldsymbol{v}^\top\boldsymbol{y}_i)^2-\mathbb{E}\Big[(\boldsymbol{w}^\top\boldsymbol{y}_1)^2(\boldsymbol{v}^\top\boldsymbol{y}_1)^2\Big]\right|\leq\epsilon;$$

These imply that for any $\boldsymbol{\theta},\boldsymbol{v}\in\mathbb{R}^p$, we have:

$$\frac{\mathbb{E}\Big[(\tilde{\boldsymbol{r}}(\boldsymbol{\theta})^\top\boldsymbol{y})^2(\tilde{\boldsymbol{w}}(\boldsymbol{\theta};\boldsymbol{v})^\top\boldsymbol{y})^2\Big]-\epsilon}{(1+\epsilon)^2}\leq g(\boldsymbol{\theta};\boldsymbol{v})\leq\frac{\mathbb{E}\Big[(\tilde{\boldsymbol{r}}(\boldsymbol{\theta})^\top\boldsymbol{y}_1)^2(\tilde{\boldsymbol{w}}(\boldsymbol{\theta};\boldsymbol{v})^\top\boldsymbol{y}_1)^2\Big]+\epsilon}{(1-\epsilon)^2}. \tag{4}$$

First, we derive the upper bound for (4):

$$\begin{aligned}
\text{RHS}=&\frac{\epsilon}{(1-\epsilon)^2}+\frac{\mathbb{E}\Big[(\tilde{\boldsymbol{r}}(\boldsymbol{\theta})^\top\boldsymbol{y})^2(\tilde{\boldsymbol{w}}(\boldsymbol{\theta};\boldsymbol{v})^\top\boldsymbol{y})^2\Big]}{(1-\epsilon)^2\Big(\tilde{\boldsymbol{r}}(\boldsymbol{\theta})^\top\tilde{\boldsymbol{r}}(\boldsymbol{\theta})\Big)\Big(\tilde{\boldsymbol{w}}(\boldsymbol{\theta};\boldsymbol{v})^\top\tilde{\boldsymbol{w}}(\boldsymbol{\theta};\boldsymbol{v})\Big)}\\
\overset{\text{Homogeneity}}{=}&\frac{\epsilon}{(1-\epsilon)^2}+\frac{\mathbb{E}\Big[((\boldsymbol{S}^{1/2}\boldsymbol{r}(\boldsymbol{\theta}))^\top\boldsymbol{y})^2((\boldsymbol{S}^{1/2}\nabla F(\boldsymbol{\theta})\boldsymbol{v})^\top\boldsymbol{y})^2\Big]}{(1-\epsilon)^2\Big(\big(\boldsymbol{S}^{1/2}\boldsymbol{r}(\boldsymbol{\theta})\big)^\top\boldsymbol{S}^{1/2}\boldsymbol{r}(\boldsymbol{\theta})\Big)\Big(\big(\boldsymbol{S}^{1/2}\nabla F(\boldsymbol{\theta})\boldsymbol{v}\big)^\top\big(\boldsymbol{S}^{1/2}\nabla F(\boldsymbol{\theta})\boldsymbol{v}\big)\Big)}\\
=&\frac{\epsilon}{(1-\epsilon)^2}+\frac{\boldsymbol{v}^\top\Sigma(\boldsymbol{\theta})\boldsymbol{v}}{2(1-\epsilon)^2\mathcal{L}(\boldsymbol{\theta})\boldsymbol{v}^\top G(\boldsymbol{\theta})\boldsymbol{v}}\overset{\text{Lemma H.1}}{=}\frac{\epsilon}{(1-\epsilon)^2}+\frac{2\mathcal{L}(\boldsymbol{\theta})\boldsymbol{v}^\top G(\boldsymbol{\theta})\boldsymbol{v}+\big(\nabla\mathcal{L}(\boldsymbol{\theta})^\top\boldsymbol{v}\big)^2}{2(1-\epsilon)^2\mathcal{L}(\boldsymbol{\theta})\boldsymbol{v}^\top G(\boldsymbol{\theta})\boldsymbol{v}}\\
=&\frac{1+\epsilon}{(1-\epsilon)^2}+\frac{\big(\nabla\mathcal{L}(\boldsymbol{\theta})^\top\boldsymbol{v}\big)^2}{2(1-\epsilon)^2\mathcal{L}(\boldsymbol{\theta})\boldsymbol{v}^\top G(\boldsymbol{\theta})\boldsymbol{v}}\overset{\text{Lemma H.2}}{\leq}\frac{1+\epsilon}{(1-\epsilon)^2}+\frac{1}{(1-\epsilon)^2}=\frac{2+\epsilon}{(1-\epsilon)^2}.
\end{aligned}$$

Moreover, if $\langle\boldsymbol{v},\mathcal{L}(\boldsymbol{\theta})\rangle=0$, then the bound is

$$\text{RHS}\leq\frac{1+\epsilon}{(1-\epsilon)^2}.$$

In the similar way, we can derive the lower bound for (4):

$$\begin{aligned}
\text{LHS}=&\frac{\boldsymbol{v}^\top\Sigma(\boldsymbol{\theta})\boldsymbol{v}}{2(1+\epsilon)^2\mathcal{L}(\boldsymbol{\theta})\boldsymbol{v}^\top G(\boldsymbol{\theta})\boldsymbol{v}}-\frac{\epsilon}{(1+\epsilon)^2}\overset{\text{Lemma H.1}}{=}\frac{2\mathcal{L}(\boldsymbol{\theta})\boldsymbol{v}^\top G(\boldsymbol{\theta})\boldsymbol{v}+\big(\nabla\mathcal{L}(\boldsymbol{\theta})^\top\boldsymbol{v}\big)^2}{2(1+\epsilon)^2\mathcal{L}(\boldsymbol{\theta})\boldsymbol{v}^\top G(\boldsymbol{\theta})\boldsymbol{v}}-\frac{\epsilon}{(1+\epsilon)^2}\\
\geq&\frac{1}{(1+\epsilon)^2}-\frac{\epsilon}{(1+\epsilon)^2}=\frac{1-\epsilon}{(1+\epsilon)^2}.
\end{aligned}$$

So for any $\boldsymbol{S}^{1/2}\boldsymbol{u}(\boldsymbol{\theta})\neq\mathbf{0},\boldsymbol{S}^{1/2}\nabla F(\boldsymbol{\theta})\boldsymbol{v}\neq0$, we have

$$\frac{1-\epsilon}{(1+\epsilon)^2}\leq g(\boldsymbol{\theta};\boldsymbol{v})\leq\frac{2+\epsilon}{(1-\epsilon)^2}.$$

Moreover, if $\langle\boldsymbol{v},\nabla\mathcal{L}(\boldsymbol{\theta})\rangle=0$, then

$$\frac{1-\epsilon}{(1+\epsilon)^2}\leq g(\boldsymbol{\theta};\boldsymbol{v})\leq\frac{1+\epsilon}{(1-\epsilon)^2}.$$

Hence, we have proved this theorem: For any $\epsilon, \delta > 0$, if $n \gtrsim \max\left\{\left(d^2 \log^2(1/\epsilon) + \log^2(1/\delta)\right)/\epsilon, \left(d \log(1/\epsilon) + \log(1/\delta)\right)/\epsilon^2\right\}$, then *w.p.* at least $1 - \delta$, the strong alignment holds uniformly:

$$\text{(i).}\ \frac{1-\epsilon}{(1+\epsilon)^2} \leq \inf_{\boldsymbol{\theta},\boldsymbol{v}\in\mathbb{R}^p} g(\boldsymbol{\theta};\boldsymbol{v}) \leq \sup_{\boldsymbol{\theta},\boldsymbol{v}\in\mathbb{R}^p} g(\boldsymbol{\theta};\boldsymbol{v}) \leq \frac{2+\epsilon}{(1-\epsilon)^2},$$

$$\text{(ii).}\ \frac{1-\epsilon}{(1+\epsilon)^2} \leq \inf_{\boldsymbol{\theta}\in\mathbb{R}^p,\langle\boldsymbol{v},\nabla\mathcal{L}(\boldsymbol{\theta})\rangle=0} g(\boldsymbol{\theta};\boldsymbol{v}) \leq \sup_{\boldsymbol{\theta}\in\mathbb{R}^p,\langle\boldsymbol{v},\nabla\mathcal{L}(\boldsymbol{\theta})\rangle=0} g(\boldsymbol{\theta};\boldsymbol{v}) \leq \frac{1+\epsilon}{(1-\epsilon)^2}.$$

□

I Proofs in Section 5: Escape directions

I.1 Proof of Theorem 5.2

Recall that $\boldsymbol{w}(t) = \sum_{i=1}^d w_i(t)\boldsymbol{u}_i$ with $w_i(t) = \boldsymbol{u}_i^\top \boldsymbol{w}(t)$. Then, $w_i(t+1) = (1-\eta\lambda_i)w_i(t) + \eta\boldsymbol{\xi}(t)^\top\boldsymbol{u}_i$. Taking the expectation of the square of both sides, we obtain

$$\mathbb{E}\left[w_i^2(t+1)\right] = (1-\eta\lambda_i)^2\mathbb{E}\left[w_i^2(t)\right] + \eta^2\mathbb{E}[|\boldsymbol{u}_i^\top\boldsymbol{\xi}(t)|^2],$$

According to Assumption 5.1, there exists $A_1, A_2 > 0$ such that for any $i \in [d]$,

$$A_1\lambda_i\mathcal{L}(\boldsymbol{w}_t) \leq \mathbb{E}[|\boldsymbol{u}_i^T\boldsymbol{\xi}(t)|] \leq A_2\lambda_i\mathcal{L}(\boldsymbol{w}_t).$$

Let $X_t = \sum_{i=1}^k \lambda_i\mathbb{E}[w_i^2(t)], Y_t = \sum_{i=k+1}^d \lambda_i\mathbb{E}[w_i^2(t)]$ denote the components of loss energy along sharp and flat directions, respectively. And we denote $D_k(t) := Y_t/X_t$.

Plugging the fact that $2\mathcal{L}(\boldsymbol{w}(t)) = X_t + Y_t$ into the two formulations above, we can obtain the following component dynamics:

$$\begin{aligned}
X_{t+1} &\leq \alpha_k X_t + A_2\eta^2(\sum_{i=1}^k \lambda_i^2)(X_t+Y_t),\\
X_{t+1} &\geq A_1\eta^2(\sum_{i=1}^k \lambda_i^2)(X_t+Y_t),\\
Y_{t+1} &\geq A_1\eta^2\big(\sum_{i=k+1}^d \lambda_i^2\big)(X_t+Y_t),
\end{aligned} \tag{5}$$

where $\alpha_k \leq \max_{i=1,\dots,k}|1-\eta\lambda_i|^2$. The terms $\alpha_k X_t$ and $\beta_k Y_t$ capture the impact of the gradient, while the remaining terms originate from the noise.

From (5), we have the following estimate about $D_k(t+1)$:

$$\begin{aligned}
D_k(t+1) = \frac{Y_{t+1}}{X_{t+1}} &\geq \frac{A_1\eta^2\left(\sum_{i=k+1}^d \lambda_i^2\right)(X_t+Y_t)}{\alpha_k X_t + A_2\eta^2(\sum_{i=1}^k\lambda_i^2)(X_t+Y_t)}\\
=&\frac{A_1\sum_{i=k+1}^d\lambda_i^2}{A_2\sum_{i=1}^k\lambda_i^2}\cdot\frac{1}{1+\frac{\alpha_k}{A_2\eta^2\sum_{i=k+1}^d\lambda_i^2}\frac{X_t}{X_t+Y_t}}\\
\geq&\frac{A_1\sum_{i=k+1}^d\lambda_i^2}{A_2\sum_{i=1}^k\lambda_i^2}\cdot\frac{1}{1+\frac{\max_{1\leq i\leq k}|1-\eta\lambda_i|^2}{A_2\eta^2\sum_{i=1}^k\lambda_i^2}\frac{X_t}{X_t+Y_t}}.
\end{aligned} \tag{6}$$

We will prove this theorem for the learning rate $\eta = \frac{\beta}{\|G(\boldsymbol{\theta})\|_F}$, where $\beta \geq \frac{1.1}{\sqrt{A_1}}$.

<u>Case (I). Small learning rate $\eta \in [\frac{1.1}{\sqrt{A_1}\|G(\boldsymbol{\theta})\|_{\mathrm{F}}}, \frac{1}{\lambda_1}]$.</u>

In this step, we consider $\eta = \frac{\beta}{\|G(\boldsymbol{\theta})\|_F}$ such that $\beta \geq \frac{1.1}{\sqrt{A_1}}$ and $\eta \leq \frac{1}{\lambda_1}$. Then we have:

$$\frac{\max\limits_{1\leq i\leq k}|1-\eta\lambda_i|^2}{A_2\eta^2\sum_{i=k+1}^d\lambda_i^2} \leq \frac{1}{A_2\eta^2\sum_{i=1}^k\lambda_i^2}.$$

Notice that (5) also ensures:

$$(X_{t+1}+Y_{t+1}) \geq A_1\eta^2\Big(\sum_{i=1}^d\lambda_i^2\Big)(X_t+Y_t).$$

Combining this inequality with (5), we have the estimate:

$$\begin{aligned}\frac{X_{t+1}}{X_{t+1}+Y_{t+1}} &\leq \frac{\alpha_k X_t + A_2\eta^2(\sum_{i=1}^k\lambda_i^2)(X_t+Y_t)}{X_{t+1}+Y_{t+1}}\\ &\leq \frac{\alpha_k X_t}{A_1\eta^2\big(\sum_{i=1}^d\lambda_i^2\big)(X_t+Y_t)} + \frac{A_2(\sum_{i=1}^k\lambda_i^2)}{A_1\big(\sum_{i=1}^d\lambda_i^2\big)}\end{aligned}$$

For simplicity, we denote $W_t := \frac{X_t}{X_t+Y_t}$, $A := \frac{\alpha_k}{A_1\eta^2\left(\sum_{i=1}^d\lambda_i^2\right)}$, and $B := \frac{A_2(\sum_{i=1}^k\lambda_i^2)}{A_1\left(\sum_{i=1}^d\lambda_i^2\right)}$.

From $\eta \leq 1/3$, we have $\alpha_k \leq 1$ and $A \leq \frac{1}{A_1\eta^2\left(\sum_{i=1}^d\lambda_i^2\right)} = \frac{1}{A_1\beta^2} < 1$. Moreover, it holds that

$$\begin{aligned}W_{t+1} &\leq AW_t + B \leq A(AW_{t-1}+B)+B = A^2W_{t-1}+B(1+A)\\ &\leq \cdots \leq A^{t+1}W_0 + B(1+A+\cdots+A^t) = A^{t+1}W_0 + \frac{1-A^{t+1}}{1-A}B\end{aligned}$$

On the one hand, if we choose

$$t \geq \frac{\log\Big(1/W_0A_2\eta^2\sum_{i=1}^k\lambda_i^2\Big)}{\log\left(A_1\beta^2\right)},$$

then we have

$$A^tW_0 \leq \left(\frac{\alpha_k}{A_1\eta^2(\sum_{i=1}^d\lambda_i^2)}\right)^t W_0 \leq \left(\frac{1}{A_1\beta^2}\right)^t W_0 \leq A_2\eta^2\sum_{i=1}^k\lambda_i^2.$$

On the other hand, if we choose $t \geq 1$, then it holds that

$$\frac{1-A^t}{1-A}B \leq B = \frac{A_2(\sum_{i=1}^k\lambda_i^2)}{A_1\big(\sum_{i=1}^d\lambda_i^2\big)} \leq A_2\eta^2\sum_{i=1}^k\lambda_i^2.$$

Hence, if we choose

$$t \geq \max\left\{1, \frac{\log\Big(1/W_0A_2\eta^2\sum_{i=1}^k\lambda_i^2\Big)}{\log\left(A_1\beta^2\right)}\right\},$$

then we have

$$\frac{X_t}{X_t+Y_t} = W_t \leq A^tW_0 + \frac{1-A^t}{1-A}B \leq 2A_2\eta^2\sum_{i=1}^k\lambda_i^2,$$

which implies that

$$\text{RHS of (6)} \geq \frac{A_1\sum_{i=k+1}^d\lambda_i^2}{A_2\sum_{i=1}^k\lambda_i^2}\cdot\frac{1}{1+\frac{\max\limits_{1\leq i\leq k}|1-\eta\lambda_i|^2}{A_2\eta^2\sum_{i=1}^k\lambda_i^2}\frac{X_t}{X_t+Y_t}}$$

$$\geq \frac{A_1 \sum_{i=k+1}^d \lambda_i^2}{A_2 \sum_{i=1}^k \lambda_i^2} \cdot \frac{1}{1 + \frac{1}{A_2 \eta^2 \sum_{i=1}^k \lambda_i^2} \cdot 2A_2 \eta^2 \sum_{i=1}^k \lambda_i^2} = \frac{A_1 \sum_{i=k+1}^d \lambda_i^2}{3A_2 \sum_{i=1}^k \lambda_i^2}.$$

Case (II). Large learning rate $\eta \geq 1/\lambda_1$.

In this step, we consider $\eta \geq \frac{1}{\lambda_1}$. Then for any $t \geq 0$, we have:

$$\text{RHS of (6)} = \frac{A_1 \sum_{i=k+1}^d \lambda_i^2}{A_2 \sum_{i=1}^k \lambda_i^2} \cdot \frac{1}{1 + \frac{\alpha_k}{\sum_{i=k+1}^d \lambda_i^2} \frac{X_t}{X_t + Y_t}} \geq \frac{A_1 \sum_{i=k+1}^d \lambda_i^2}{A_2 \sum_{i=1}^k \lambda_i^2} \cdot \frac{1}{1 + \frac{\max_{i \in [k]} |1-\eta\lambda_i|^2}{A_2 \eta^2 \sum_{i=1}^k \lambda_i^2}}$$

$$\geq \frac{A_1 \sum_{i=k+1}^d \lambda_i^2}{A_2 \sum_{i=1}^k \lambda_i^2} \cdot \frac{1}{1 + \frac{\max\{1, |1-\eta\lambda_1|^2\}}{A_2 \eta^2 \sum_{i=1}^k \lambda_i^2}} \geq \frac{A_1 \sum_{i=k+1}^d \lambda_i^2}{A_2 \sum_{i=1}^k \lambda_i^2} \cdot \frac{1}{1 + \frac{1}{A_2}} = \frac{A_1 \sum_{i=k+1}^d \lambda_i^2}{(A_2 + 1) \sum_{i=1}^k \lambda_i^2}.$$

Combining Case (I) and (II), we obtain this theorem: If we choose the learning rate $\eta = \frac{\beta}{\|G(\boldsymbol{\theta})\|_F}$, where $\beta \geq \frac{1.1}{\sqrt{A_1}}$, then for any

$$t \geq \max\left\{1, \frac{\log\left(1/W_0 A_2 \eta^2 \sum_{i=1}^k \lambda_i^2\right)}{\log(A_1 \beta^2)}\right\},$$

we have

$$D_k(t+1) \geq \frac{A_1 \sum_{i=k+1}^d \lambda_i^2}{\max\{3A_2, A_2 + 1\} \sum_{i=1}^k \lambda_i^2}.$$

□

I.2 Proof of Proposition 5.3

Recall that $\boldsymbol{w}(t) = \sum_{i=1}^d w_i(t) \boldsymbol{u}_i$ with $w_i(t) = \boldsymbol{u}_i^\top \boldsymbol{w}(t)$. Then, for GD, $w_i(t+1) = (1 - \eta\lambda_i) w_i(t)$, which implies:

$$w_i(t) = (1 - \eta\lambda_i)^t w_i(0).$$

Therefore, for $\eta = \beta/\lambda_1$ ($\beta > 2$), it holds that

$$D_1(t) = \frac{\sum_{i=2}^d \lambda_i w_i^2(t)}{\lambda_1 w_1^2(t)} = \frac{\sum_{i=2}^d \lambda_i (1-\eta\lambda_i)^{2t} w_i^2(0)}{\lambda_1 (1-\eta\lambda_1)^{2t} w_1^2(0)}.$$

□

J Useful Inequalities

Lemma J.1 (Bernstein's Inequality (Vershynin, 2018))**.** *Suppose $\{X_1, \cdots, X_n\}$ are independent sub-Exponential random variables with $\|X_i\|_{\psi_1} \leq K$. Then there exists an absolute constant $c > 0$ such that for any $t \geq 0$, we have:*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_i - \frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i]\right| > t\right) \leq 2\exp\left(-cn \min\left\{\frac{t}{K}, \frac{t^2}{K^2}\right\}\right).$$

Lemma J.2 (Hanson-Wright's Inequality (Vershynin, 2018))**.** *Let $\boldsymbol{X} = (X_1, \cdots, X_n) \in \mathbb{R}^n$ be a random vector with independent mean zero sub-Gaussian coordinates. Let $\boldsymbol{A}$ be an $n \times n$ matrix. Then, there exists an absolute constant c such that for every $t \geq 0$, we have*

$$\mathbb{P}\left(\left|\boldsymbol{X}^\top \boldsymbol{A} \boldsymbol{X} - \mathbb{E}[\boldsymbol{X}^\top \boldsymbol{A} \boldsymbol{X}]\right| \geq t\right) \leq 2\exp\left(-c \min\left\{\frac{t^2}{K^4 \|\boldsymbol{A}\|_F^2}, \frac{t}{K^2 \|\boldsymbol{A}\|_2}\right\}\right),$$

where $K = \max_i \|X_i\|_{\psi_2}$.

Lemma J.3 (Covariance Estimate for sub-Gaussian Distribution (Vershynin, 2018))**.** *Let $\boldsymbol{x}, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$ be i.i.d. random vectors in $\mathbb{R}^d$. More precisely, assume that there exists $K \geq 1$ s.t. $\|\langle \boldsymbol{x}, \boldsymbol{v}\rangle\|_{\psi_2} \leq K \|\langle \boldsymbol{x}, \boldsymbol{v}\rangle\|_{L_2}$ for any $\boldsymbol{v} \in \mathbb{S}^{d-1}$, Then for any $u \geq 0$,* w.p. *at least $1 - 2\exp(-u)$ one has*

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top - \mathbb{E}\big[\boldsymbol{x}\boldsymbol{x}^\top\big] \right\| \leq CK^2 \left(\sqrt{\frac{d+u}{n}} + \frac{d+u}{n} \right) \left\| \mathbb{E}\big[\boldsymbol{x}\boldsymbol{x}^\top\big] \right\|,$$

where C is an absolute positive constant.

Definition J.4 (Sub-Weibull Distribution)**.** We define X as a sub-Weibull random variable if it has a bounded ψ_β-norm. The ψ_β-norm of X for any $\beta > 0$ is defined as

$$\|X\|_{\psi_\beta} := \inf \Big\{ C > 0 : \mathbb{E}\big[\exp(|X|^\beta / C^\beta)\big] \leq 2 \Big\}.$$

Particularly, when $\beta = 1$ or 2, sub-Weibull random variables reduce to sub-Exponential or sub-Gaussian random variables, respectively.

Lemma J.5 (Concentration Inequality for Sub-Weibull Distribution, Theorem 3.1 in (Hao et al., 2019))**.** *Suppose $\{X_i\}_{i=1}^n$ are independent sub-Weibull random variables with $\|X_i\|_{\psi_\beta} \leq K$. Then there exists an absolute constant C_β only depending on β such that for any $\delta \in (0, 1/e^2)$,* w.p. *at least $1 - \delta$, we have*

$$\left| \frac{1}{n} \sum_{i=1}^{n} X_i - \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\big[X_i\big] \right| \leq C_\beta K \left(\Big(\frac{\log(1/\delta)}{n} \Big)^{1/2} + \frac{\big(\log(1/\delta)\big)^{1/\beta}}{n} \right).$$

Lemma J.6 (Cauchy-Schwarz Inequalities)**.**
(1) Let $S \in \mathbb{R}^{n \times n}$ be a positive symmetric definite matrix. For any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, we denote $\langle \boldsymbol{x}, \boldsymbol{y} \rangle_S := \boldsymbol{x}^\top S \boldsymbol{y}$ and $\|\boldsymbol{x}\|_S := \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle_S}$, then we have $|\langle \boldsymbol{x}, \boldsymbol{y} \rangle_S| \leq \|\boldsymbol{x}\|_S \|\boldsymbol{y}\|_S$.
(2) Given two random variables X and Y, it holds that $|\mathbb{E}[XY]| \leq \sqrt{\mathbb{E}[X^2]}\sqrt{\mathbb{E}[Y^2]}$.