Improved Scaling Laws in Linear Regression via Data Reuse

Licong Lin
UC Berkeley
liconglin@berkeley.edu

Jingfeng Wu
UC Berkeley
uuujf@berkeley.edu

Peter L. Bartlett
UC Berkeley and Google DeepMind
peter@berkeley.edu

Abstract

Neural scaling laws suggest that the test error of large language models trained online decreases polynomially as the model size and data size increase. However, such scaling can be unsustainable when running out of new data. In this work, we show that data reuse can improve existing scaling laws in linear regression. Specifically, we derive sharp test error bounds on M-dimensional linear models trained by multi-pass *stochastic gradient descent* (multi-pass SGD) on N data with sketched features. Assuming that the data covariance has a power-law spectrum of degree a, and that the true parameter follows a prior with an aligned power-law spectrum of degree $b-a$ (with $a > b > 1$), we show that multi-pass SGD achieves a test error of $\Theta(M^{1-b} + L^{(1-b)/a})$, where $L \lesssim N^{a/b}$ is the number of iterations. In the same setting, one-pass SGD only attains a test error of $\Theta(M^{1-b} + N^{(1-b)/a})$ (see, e.g., Lin et al., 2024). This suggests an improved scaling law via data reuse (i.e., choosing $L > N$) in data-constrained regimes. Numerical simulations are also provided to verify our theoretical findings.

1 Introduction

Empirical studies reveal that the performance of large-scale models often improves in a predictable manner as both model size (denoted by M) and sample size (denoted by N) increase (see, e.g., Hoffmann et al., 2022; Besiroglu et al., 2024). These observations, known as *neural scaling laws*, suggest that the population risk (denoted by $\mathcal{R}$) of large models decreases following a power-law formula, namely,

$$\mathcal{R}(M, N) \approx \mathcal{R}^* + c_1 M^{-a_1} + c_2 N^{-a_2},$$

where $\mathcal{R}^* > 0$ denotes the irreducible error—such as the intrinsic entropy of natural language in the case of language modeling (Kaplan et al., 2020)—and a_1, a_2, c_1, c_2 are positive constants. Neural scaling laws predict a path for improving the state-of-the-art models via *scaling model and data size*.

A line of recent work establishes provable scaling laws in simplified settings such as linear regression (see, e.g., Lin et al., 2024; Paquette et al., 2024, other related works will be discussed later in Section 6). Specifically, they consider an infinite-dimensional linear regression problem, where an M-dimensional linear model is trained by one-pass *stochastic gradient descent* (SGD) on N Gaussian-sketched samples. Under power-law assumptions on the spectra of the data covariance and the prior covariance, they show power-law type scaling laws in linear regression. However, their results are limited to one-pass SGD, where each sample is used once. In particular, Lin et al. (2024) attributed the nice, power-law type scaling laws to the *implicit regularization* effect of one-pass SGD

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

(see Section 1 therein). It is unclear if scaling laws apply to other training algorithms, particularly those involving multiple passes of the data. Indeed, the scaling laws that only apply to one-pass methods are not sustainable in a data-constrained regime.

There is evidence that data reuse can improve existing scaling laws developed for one-pass training. Empirically, Muennighoff et al. (2023) showed that with up to four passes, scaling laws approximately hold as if the reused data is new. From a theoretical perspective, the work by Pillaud-Vivien et al. (2018) shows that in a class of linear regression problems, the sample complexity of one-pass SGD is strictly suboptimal; and it can be made minimax optimal by considering multiple passes. However, Pillaud-Vivien et al. (2018) did not discuss the effect of model size or sketching. These results motivate the study of scaling laws for multi-pass methods.

Contributions. In this work, we study scaling laws induced by multi-pass SGD in the same infinite-dimensional linear regression setting considered by Lin et al. (2024); Paquette et al. (2024). Our results suggest that in certain regimes, the test error of models trained by multi-pass SGD scales strictly better with respect to the number of training samples compared to one-pass SGD.

We assume that the data covariance and the prior covariance exhibit aligned power-law spectra with exponents a and $b-a$, respectively (see Assumption 1C and 1D) (Lin et al., 2024; Paquette et al., 2024). We prove that multi-pass SGD achieves an excess test error of order $\Theta(M^{1-b}+L^{(1-b)/a})$ when $a > b > 1$ and the number of SGD iterations $L \lesssim N^{a/b}$. This improves over the $\Theta(M^{1-b}+N^{(1-b)/a})$ bound for one-pass SGD (Lin et al., 2024) when $L > N$. In particular, when choosing the optimal number of iterations $L \eqsim N^{a/b}$, multi-pass SGD achieves an excess test error of order $\Theta(N^{(1-b)/b})$, in contrast to $\Theta(N^{(1-b)/a})$ for one-pass SGD in the data-constrained regime where $N \ll M^b$. Our results thus suggest that, to a certain extent, reusing data can improve the test performance of linear models in data-constrained regimes.

Notation. Let $f(x)$ and $g(x)$ be two positive-valued functions. We write $f(x) \lesssim g(x)$ (and $f(x) = \mathcal{O}(g(x))$) if there exists some absolute constant (if not otherwise specified) $c > 0$ such that $f(x) \leqslant cg(x)$ for all x. Similarly, $f(x) \gtrsim g(x)$ (and $f(x) = \Omega(g(x))$) means $f(x) \geqslant cg(x)$ for some constant $c > 0$. We write $f(x) \eqsim g(x)$ (and $f(x) = \Theta(g(x))$) when $f(x) \lesssim g(x) \lesssim f(x)$. We also occasionally use $\widetilde{\mathcal{O}}(\cdot), \widetilde{\Theta}(\cdot)$ to hide logarithmic factors. In this work, $\log(\cdot)$ denotes the base-2 logarithm. For two vectors $\mathbf{u}, \mathbf{v}$ in a Hilbert space, we denote their inner product by $\langle \mathbf{u}, \mathbf{v} \rangle$ or $\mathbf{u}^\top \mathbf{v}$. We denote the operator norm for matrices by $\|\cdot\|$ (or $\|\cdot\|_2$) and the ℓ_2-norm for vectors by $\|\cdot\|_2$. For a positive semi-definite (PSD) matrix $\boldsymbol{A}$ and a vector $\boldsymbol{v}$ of compatible dimensions, we write $\|\boldsymbol{v}\|_{\boldsymbol{A}}^2 := \boldsymbol{v}^\top \boldsymbol{A} \boldsymbol{v}$. For symmetric matrices, we denote the j-th eigenvalue of $\boldsymbol{A}$ by $\mu_j(\boldsymbol{A})$, and the rank of $\boldsymbol{A}$ by $r(\boldsymbol{A})$.

2 Setup

Let $\mathbf{x} \in \mathbb{H}$ denote a feature vector in a Hilbert space $\mathbb{H}$ (finite or countably infinite-dimensional) with dimension $d := \dim(\mathbb{H})$, and $y \in \mathbb{R}$ denote its corresponding response. In linear regression, the test error (i.e., population risk) of the parameter $\mathbf{w} \in \mathbb{H}$ is measured by the mean squared error:

$$\mathcal{R}(\mathbf{w}) := \mathbb{E}_{(\mathbf{x},y)\sim P}\Big[\big(\langle \mathbf{x}, \mathbf{w} \rangle - y\big)^2\Big]$$

for some distribution P on $\mathbb{H} \times \mathbb{R}$. Given samples of the form $(\mathbf{x}, y)$, instead of fitting a d-dimensional linear model, we train an M-dimensional sketched linear model with $M \ll d$. Namely, we consider linear predictors with M parameters, defined as

$$f_{\mathbf{v}} : \mathbb{H} \to \mathbb{R}, \quad \mathbf{x} \mapsto \langle \mathbf{v}, \mathbf{S}\mathbf{x} \rangle, \tag{1}$$

where $\mathbf{v} \in \mathbb{R}^M$ are the trainable parameters, and $\mathbf{S} \in \mathbb{R}^{M \times d}$ is some fixed sketching matrix. In this work, we consider Gaussian sketching, where the entries of $\mathbf{S}$ are drawn independently from $\mathcal{N}(0, 1/M)$. Given a set of N i.i.d. samples $(\mathbf{x}_i, y_i)_{i=1}^N$ from P, we train $f_{\mathbf{v}}$ via *multi-pass stochastic gradient descent* (multi-pass SGD), that is,

$$\begin{aligned} \mathbf{v}_t &:= \mathbf{v}_{t-1} - \gamma_t \big(f_{\mathbf{v}_{t-1}}(\mathbf{x}_{i_t}) - y_{i_t}\big) \nabla_{\mathbf{v}} f_{\mathbf{v}_{t-1}}(\mathbf{x}_{i_t}) \\ &= \mathbf{v}_{t-1} - \gamma_t \mathbf{S}\mathbf{x}_{i_t}(\mathbf{x}_{i_t}^\top \mathbf{S}^\top \mathbf{v}_{t-1} - y_{i_t}), \qquad t = 1, \dots, L, \end{aligned} \tag{multi-pass SGD}$$

where L is the number of total steps, $i_t \overset{iid}{\sim} \mathsf{unif}([N])$ for $t \in [L]$, and $(\gamma_t)_{t=1}^L$ are the stepsizes. Without loss of generality, we assume zero initialization $\mathbf{v}_0 = \mathbf{0}$. We consider a geometric decaying stepsize scheduler (Ge et al., 2019; Wu et al., 2022b; Lin et al., 2024),

$$\gamma_t := \gamma_0/2^\ell \quad \text{for } t = 1, \dots, L, \quad \text{where } \ell = \lfloor t/(L/\log(L)) \rfloor, \tag{2}$$

and $\gamma_0 > 0$ is the initial stepsize. The output of multi-pass SGD is taken as its last iterate $\mathbf{v}_L$. We emphasize that the algorithm we consider differs slightly from the standard SGD used in practice, where the samples are shuffled at the beginning of each epoch (pass) and then processed sequentially without replacement. In contrast, we assume that at each step, a sample is drawn independently from the training dataset, allowing for repeated sampling within an epoch. Moreover, our analysis applies to other stepsize schedules (such as polynomial decay), but we focus on geometric decay since it is known to yield near minimax optimal excess test error for the last iterate of SGD in the finite-dimensional regime (Ge et al., 2019).

Conditioned on a sketching matrix $\mathbf{S}$, the risk of $\mathbf{v}_L$ is computed as

$$\mathcal{R}_M(\mathbf{v}_L) = \mathcal{R}(\mathbf{S}^\top \mathbf{v}_L) = \mathbb{E}\Big[\big(\langle \mathbf{x}, \mathbf{S}^\top \mathbf{v}_L \rangle - y\big)^2\Big],$$

where the expectation is over $(\mathbf{x}, y)$ from P. As an important component of our analysis, we also consider the *gradient descent* (GD) iterates

$$\begin{aligned} \boldsymbol{\theta}_t &:= \boldsymbol{\theta}_{t-1} - \frac{\gamma_t}{N} \sum_{i=1}^N \big(f_{\boldsymbol{\theta}_{t-1}}(\mathbf{x}_i) - y_i\big) \nabla_{\mathbf{v}} f_{\boldsymbol{\theta}_{t-1}}(\mathbf{x}_i) \\ &= \boldsymbol{\theta}_{t-1} - \frac{\gamma_t}{N} \mathbf{S}\mathbf{X}^\top (\mathbf{X}\mathbf{S}^\top \boldsymbol{\theta}_{t-1} - \mathbf{y}), \qquad\qquad t = 1, \dots, L, \end{aligned} \tag{GD}$$

where $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)^\top$, $\mathbf{y} = (y_1, \dots, y_N)^\top$, $\boldsymbol{\theta}_0 = \mathbf{0}$, and $(\gamma_t)_{t=1}^L$ are the same stepsizes as in (2). Conditioned on the sketching matrix $\mathbf{S}$ and the dataset $(\mathbf{x}_i, y_i)_{i=1}^N$, it can be verified by induction that $\mathbf{v}_L$ is an unbiased estimate of $\boldsymbol{\theta}_L$, i.e., $\mathbb{E}[\mathbf{v}_L] = \boldsymbol{\theta}_L$, where the expectation is over the randomness of the indices $(i_t)_{t=1}^L$.

Risk decomposition. We can decompose the risk (i.e., the test error) achieved by $\mathbf{v}_L$, the last iterate of (multi-pass SGD), into the sum of *irreducible risk*, *approximation error*, the *excess risk* of the last iterate of (GD), and a *fluctuation error*:

$$\mathcal{R}_M(\mathbf{v}_L) = \underbrace{\min \mathcal{R}(\cdot)}_{\mathsf{Irreducible}} + \underbrace{\min \mathcal{R}_M(\cdot) - \min \mathcal{R}(\cdot)}_{\mathsf{Approx}} + \underbrace{\mathcal{R}_M(\boldsymbol{\theta}_L) - \min \mathcal{R}_M(\cdot)}_{\mathsf{Excess}} + \underbrace{\mathcal{R}_M(\mathbf{v}_L) - \mathcal{R}_M(\boldsymbol{\theta}_L)}_{\mathsf{Fluc}}. \tag{3}$$

Compared with Lin et al. (2024) (cf. Eq. 4), the decomposition in (3) includes an additional *fluctuation error* term arising from the randomness of the indices $(i_t)_{t=1}^L$ in multi-pass SGD (Zou et al., 2022). Note that the fluctuation error is non-negative by Jensen's inequality, as $\mathbf{v}_L$ is an unbiased estimate of $\boldsymbol{\theta}_L$.

3 Main results

In this section, we present our main result, showing that under certain power-law assumptions on the data covariance and the prior covariance, the expected risk of $\mathbf{v}_L$ from (multi-pass SGD) decays polynomially in the number of training steps L and model size M. We begin by introducing the data assumption used throughout this work.

Assumption 1. *Assume the following conditions on the data distribution P.*

A. ***Gaussian design.*** *The feature vector satisfies $\mathbf{x} \sim \mathcal{N}(0, \mathbf{H})$.*

B. ***Well-specified model.*** *The response satisfies $\mathbb{E}[y \mid \mathbf{x}, \mathbf{w}^*] = \mathbf{x}^\top \mathbf{w}^*$. Define $\sigma^2 := \mathbb{E}[(y - \mathbf{x}^\top \mathbf{w}^*)^2]$.*

C. ***Power-law spectrum.*** *The eigenvalues of $\mathbf{H}$ satisfy $\lambda_i \eqsim i^{-a}$ for all $i > 0$ for some $a > 1$.*

D. ***Source condition.*** *Let $(\lambda_i, \mathbf{v}_i)_{i>0}$ be the eigenvalues and eigenvectors of $\mathbf{H}$. Assume $\mathbf{w}^*$ follows a prior such that*

for $i \neq j$, $\mathbb{E}[\langle \mathbf{v}_i, \mathbf{w}^* \rangle \langle \mathbf{v}_j, \mathbf{w}^* \rangle] = 0$; and for $i > 0$, $\mathbb{E}[\lambda_i \langle \mathbf{v}_i, \mathbf{w}^* \rangle^2] \eqsim i^{-b}$, for some $b > 1$.

Assumptions 1A and 1B posit that the feature vector $\mathbf{x}$ follows a Gaussian distribution and that the linear model $y = \langle \mathbf{x}, \mathbf{w}^* \rangle + \epsilon$ is well-specified, which are standard conditions in the analysis of linear regression. Assumptions 1C and 1D assume that both the covariance of $\mathbf{x}$ and the prior on the true parameter $\mathbf{w}^*$ have power-law spectra and share the same eigenspace. In particular, the true parameter $\mathbf{w}^*$ follows an isotropic prior when $a = b$. These conditions are common in theoretical analysis of scaling laws (Bordelon et al., 2024a; Lin et al., 2024; Paquette et al., 2024), and the power-law spectrum in Assumption 1C is also empirically observed in real-world data, such as the frequency distribution of words in natural languages (Piantadosi, 2014). We further note that Assumption 1 matches the conditions of Theorem 4.2 in Lin et al. (2024), which established scaling laws for one-pass SGD under the same setup. This alignment allows a direct comparison between our results and those in Lin et al. (2024), highlighting the benefits of data reuse in certain data-constrained regimes. Finally, we define the number of effective steps $L_{\mathtt{eff}} := \lfloor L/\log L \rfloor$.

Theorem 3.1 (Error bounds for multi-pass SGD)**.** *Suppose Assumption 1 holds. Consider an M-dimensional linear predictor trained by* (multi-pass SGD) *on N samples. Recall the decomposition in* (3)*. Assume the initial stepsize $\gamma_0 = \min\{\gamma, 1/[4\max_i \|\mathbf{S}\mathbf{x}_i\|_2^2]\}$ for some $\gamma \lesssim 1/\log N$ and the number of effective steps $L_{\mathtt{eff}} \lesssim N^a/\gamma$. Then with probability at least $1 - e^{-\Omega(M)}$ over $\mathbf{S}$*

1. $\mathsf{Irreducible} = \mathcal{R}(\mathbf{w}^*) = \sigma^2$.
2. $\mathbb{E}_{\mathbf{w}^*}[\mathsf{Approx}] \eqsim M^{1-b}$.
3. *Suppose $\sigma^2 \gtrsim 1$. The expected excess risk (*Excess*) admits a decomposition into a bias term (*Bias*) and a variance term (*Var*), namely,*

$$\mathbb{E}[\mathsf{Excess}] \eqsim \mathsf{Bias} + \sigma^2 \mathsf{Var},$$

where the expectation is over the randomness of $\mathbf{w}^*$, $(\mathbf{x}_i, y_i)_{i=1}^N$ and $(i_t)_{t=1}^L$. Moreover, when $a > b - 1$, Bias *and* Var *satisfy*

$$\begin{aligned}
&\mathsf{Bias} \lesssim \max\{M^{1-b}, (L_{\mathtt{eff}}\gamma)^{(1-b)/a}\},\\
&\mathsf{Bias} \gtrsim (L_{\mathtt{eff}}\gamma)^{(1-b)/a} \text{ when } (L_{\mathtt{eff}}\gamma)^{1/a} \leqslant M/c \text{ for some constant } c > 0,\\
&\mathsf{Var} \eqsim \min\left\{M,\ (L_{\mathtt{eff}}\gamma)^{1/a}\right\}/N.
\end{aligned}$$

4. *Suppose $\sigma^2 \eqsim 1$ and $L_{\mathtt{eff}} \lesssim N^{(1-\varepsilon)a}/\gamma$ for some $\varepsilon \in (0, 1]$. The expected fluctuation error $\mathbb{E}[\mathsf{Fluc}]$ satisfies*

$$\begin{aligned}
&\mathbb{E}[\mathsf{Fluc}] \lesssim \gamma \log N \cdot \big[(L_{\mathtt{eff}}\gamma)^{1/a-1} + \frac{(L_{\mathtt{eff}}\gamma)^{1/a}}{N}\big], \quad \text{and}\\
&\mathbb{E}[\mathsf{Fluc}] \gtrsim \gamma (L_{\mathtt{eff}}\gamma)^{1/a-1} \quad \text{when } L_{\mathtt{eff}} \lesssim N/\gamma \text{ and } (L_{\mathtt{eff}}\gamma)^{1/a} \leqslant M/c \text{ for some constant } c > 0,
\end{aligned}$$

where the expectation is over $\mathbf{w}^*, (\mathbf{x}_i, y_i)_{i=1}^N$ and $(i_t)_{t=1}^L$.

In the results, the hidden constants depend only on (a, b) for part 1—3, and on (a, b, ε) for part 4.

See the proof of Theorem 3.1 in Appendix A.2.1. A few comments on Theorem 3.1 are in order.

Comparison with Lin et al. (2024). The results in Theorem 3.1 are more general than those in Theorem 4.1 and 4.2 of Lin et al. (2024). Specifically, we derive matching upper and lower bounds for each term in the decomposition (3) for multi-pass SGD with an arbitrary number of steps $L \lesssim N^a$, except for the lower bound on the fluctuation error, which requires $L \lesssim N$. In contrast, Lin et al. (2024) only considered one-pass SGD where $L = N$. When $a \geqslant b$ and $L = N$, our bounds match those for one-pass SGD given in Theorems 4.1 and 4.2 of Lin et al. (2024) up to logarithmic factors (see Section 3.2 for more details).

The fluctuation error. From part 4 of Theorem 3.1, we see that the fluctuation error term $\mathbb{E}[\mathsf{Fluc}]$ vanishes as the stepsize γ goes to zero. This is intuitive: when γ is small, the noise from random sampling becomes negligible and multi-pass SGD closely approximates gradient descent. Moreover, when $a \geqslant b$ and $L_{\mathtt{eff}} \lesssim N^{a/b}$, it can be verified that for any $\gamma \lesssim 1/\log N$, the fluctuation error is dominated by the sum of the approximation error and excess risk of (GD), i.e., $\mathbb{E}[\mathsf{Fluc}] \lesssim \mathbb{E}_{\mathbf{w}^*}[\mathsf{Approx}] + \mathbb{E}[\mathsf{Excess}]$.

Choice of the stepsize. The assumption $\gamma \lesssim 1/\log N$ ensures that the initial stepsize $\gamma_0 = \gamma$ with high probability. However, to guarantee the convergence of GD iterates, it suffices to have $\gamma_0 \approx \gamma \leqslant 1/\|\mathbf{S}\mathbf{X}^\top\mathbf{X}\mathbf{S}^\top/N\|_2 \overset{(i)}{\eqsim} 1$.[1] The additional $\log N$ factor and the assumption $\gamma_0 = \min\{\gamma, 1/[4\max_i \|\mathbf{S}\mathbf{x}_i\|_2^2]\}$ are technical conditions needed for analyzing the fluctuation error term. We leave the problem of relaxing these assumptions to future work.

Constant-stepsize SGD with iterate averaging. Similar to Lin et al. (2024), the results in Theorem 3.1 also hold for the average of the iterates of multi-pass SGD with a constant stepsize, with the only modification being that $L_{\texttt{eff}}$ is replaced by L in the bounds. We provide simulations supporting this claim in Section 4.

Relaxation of Assumption 1. The Gaussian design in Assumption 1A can be relaxed to a sub-Gaussian design when establishing the upper bounds for Bias, Var, Approx in Theorem 3.1 and the upper bounds in subsequent corollaries. Moreover, the exact alignment of the eigenvectors of the prior and data covariance in Assumption 1D can be relaxed. We refer to Appendix A.3 for more details.

Next, we discuss some implications of the error bounds in Theorem 3.1.

3.1 Scaling laws for GD

To begin with, we present matching upper and lower bounds for the expected test error of the last iterate of (GD) (denoted by $\mathbb{E}[\mathcal{R}_M(\boldsymbol{\theta}_L)]$). We note that the GD iterates have strictly smaller test error than the corresponding multi-pass SGD iterates when $\gamma > 0$, since the GD iterates $(\boldsymbol{\theta}_t)_{t=1}^L$ are the expectation of the multi-pass SGD iterates $(\mathbf{v}_t)_{t=1}^L$, conditioned on the sketching matrix $\mathbf{S}$ and the dataset $(\mathbf{x}_i, y_i)_{i=1}^N$. By combining part 1—3 of Theorem 3.1, we have

Corollary 3.2 (Scaling laws for GD)**.** *Let Assumption 1 hold and $a > b - 1$. Consider an M-dimensional linear predictor trained by* (GD) *on N samples with stepsizes $\gamma_0 = \min\{\gamma, 1/[4\,\mathrm{tr}(\mathbf{S}\mathbf{X}^\top\mathbf{X}\mathbf{S}^\top/N)]\}$ for some $\gamma \lesssim 1$. Suppose $\sigma^2 \eqsim 1$ and $L_{\texttt{eff}} \lesssim N^a/\gamma$. With probability at least $1 - e^{-\Omega(M)}$ over $\mathbf{S}$, the expected risk of $\boldsymbol{\theta}_L$ satisfies*

$$\mathbb{E}[\mathcal{R}_M(\boldsymbol{\theta}_L)] = \sigma^2 + \underbrace{\Theta\Big(\frac{1}{M^{b-1}}\Big) + \Theta\Big(\frac{1}{(L_{\texttt{eff}}\gamma)^{(b-1)/a}}\Big)}_{\mathsf{Approx}+\mathsf{Bias}} + \underbrace{\Theta\Big(\frac{\min\{M, (L_{\texttt{eff}}\gamma)^{1/a}\}}{N}\Big)}_{\mathsf{Var}}.$$

Here, $\Theta(\cdot)$ hides constants that only depend on (a, b).

See the proof of Corollary 3.2 in Appendix A.2.2. From Corollary 3.2, we see that the variance error of GD is dominated by the sum of the approximation error and the bias error (i.e. $\mathsf{Var} \lesssim \mathsf{Approx} + \mathsf{Bias}$) when $L_{\texttt{eff}}\gamma \lesssim N^{a/b}$. To achieve the optimal expected test error, we may choose $\gamma \eqsim 1$ and the number of effective steps $L_{\texttt{eff}} \eqsim \min\{N^{a/b}, M^a\}/\gamma \lesssim N^a$. Under this choice, we have

$$\mathbb{E}[\mathcal{R}_M(\boldsymbol{\theta}_L)] - \sigma^2 = \begin{cases} \Theta\Big(\frac{1}{N^{(b-1)/b}}\Big), & \text{if } N \lesssim M^b, \\ \Theta\Big(\frac{1}{M^{b-1}}\Big), & \text{if } N \gtrsim M^b. \end{cases}$$

It is worth mentioning that a decreasing stepsize schedule as in (2) is not necessary for our analysis. In fact, Corollary 3.2 remains valid for the last iterate of constant-stepsize GD (i.e., $\gamma_t \equiv \gamma$) when replacing $L_{\texttt{eff}}$ with L in the bounds. In addition, the GD iterate $\boldsymbol{\theta}_L$ achieves the same expected risk (up to logarithmic factors) as one-pass SGD when $L \eqsim N$, where the performance of one-pass SGD is characterized in Theorem 4.2 of Lin et al. (2024).

However, the computational cost of GD is substantially larger than that of one-pass SGD, since each update requires computing gradients from all samples, resulting in a complexity of $\widetilde{\mathcal{O}}(MN^2)$ compared to $\widetilde{\mathcal{O}}(MN)$ for one-pass SGD. Nevertheless, the excess test error of GD serves as an always-valid lower bound for that of multi-pass SGD, and is also an upper bound (up to logarithmic factors) in certain regimes where the fluctuation error is dominated by the sum of the approximation error and the excess risk of GD.

[1] Step (i) follows from e.g., Theorem 4 and 5 in Koltchinskii and Lounici (2017).

3.2 Scaling laws for multi-pass SGD

We now analyze the expected test error of the last iterate of (multi-pass SGD). By Theorem 3.1 and Corollary 3.2, we have

Corollary 3.3 (Scaling laws for multi-pass SGD when $a \geqslant b$)**.** *Suppose the assumptions in Theorem 3.1 are in force, $\sigma^2 \eqsim 1$, and $a \geqslant b > 1$. For any $L_{\texttt{eff}} \lesssim N^{a/b}/\gamma$, we have*

$$\mathbb{E}[\mathcal{R}_M(\mathbf{v}_L)] = \sigma^2 + \Theta\bigg(\frac{1}{M^{b-1}}\bigg) + \Theta\bigg(\frac{1}{(L_{\texttt{eff}}\gamma)^{(b-1)/a}}\bigg)$$

with probability at least $1 - e^{-\Omega(M)}$. Here, all hidden constants depend only on (a, b).

In contrast, Theorem 4.2 in Lin et al. (2024) proved that one-pass SGD with $\mathbf{v}_0^o = \mathbf{0}$, $\mathbf{v}_t^o = \mathbf{v}_{t-1}^o - \gamma_t \mathbf{S}\mathbf{x}_t(\mathbf{x}_t^\top \mathbf{S}^\top \mathbf{v}_{t-1}^o - y_t)$ for $t \in [N]$ satisfies

$$\mathbb{E}[\mathcal{R}_M(\mathbf{v}_N^o)] = \sigma^2 + \Theta\bigg(\frac{1}{M^{b-1}}\bigg) + \Theta\bigg(\frac{1}{(N_{\texttt{eff}}\gamma)^{(b-1)/a}}\bigg)$$

with probability at least $1 - e^{-\Omega(M)}$, where $N_{\texttt{eff}} := N/\log N$.

Several remarks on Corollary 3.3 are listed below.

Benefits of data reuse. When $a \geqslant b > 1$, Corollary 3.3 shows that multi-pass SGD achieves an excess test error of order $\Theta(M^{1-b} + (L_{\texttt{eff}}\gamma)^{(1-b)/a})$ when the number of effective SGD steps $L_{\texttt{eff}} \lesssim N^{a/b}$, while one-pass SGD achieves an excess test error of order $\Theta(M^{1-b} + (N_{\texttt{eff}}\gamma)^{(1-b)/a})$. Therefore, the reused data across multiple passes (epochs) can be viewed as fresh data when the number of passes is smaller than $N^{a/b-1}$. For example, when $L = kN$ for some constant $k > 1$, the test error achieved by k-pass SGD matches that of one-pass SGD trained on kN $i.i.d.$ samples despite the training data being reused—aligning with the empirical observations in Muennighoff et al. (2023).

Moreover, when the number of effective steps is chosen[2] as $L_{\texttt{eff}} \eqsim \min\{N^{a/b}, M^a\}/\gamma$ and the learning rate $\gamma \eqsim 1/\log N$, the excess test error of multi-pass SGD satisfies

$$\mathbb{E}[\mathcal{R}_M(\mathbf{v}_L)] - \sigma^2 \eqsim M^{1-b} + N^{(1-b)/b},$$

while choosing $\gamma \eqsim 1$ for one-pass SGD yields

$$\mathbb{E}[\mathcal{R}_M(\mathbf{v}_N^o)] - \sigma^2 \eqsim M^{1-b} + N_{\texttt{eff}}{}^{(1-b)/a}.$$

Therefore, in the data-constrained regime where $N \ll M^b$, reusing data and running multi-pass SGD for $N^{a/b-1}$ epochs yields an improved rate of $\widetilde{\mathcal{O}}(N^{(1-b)/b})$ compared to the one-pass SGD rate of $\widetilde{\mathcal{O}}(N^{(1-b)/a})$ when $a > b$.

Optimal compute allocation. Given a total compute budget $C = L \cdot M$, by Corollary 3.3, we can set $L = \mathcal{O}(C^{a/(a+1)})$ and $M = \mathcal{O}(C^{1/(a+1)})$ with stepsize $\gamma \eqsim 1/\log L$ to achieve the optimal rate $\widetilde{\mathcal{O}}(C^{(1-b)/(a+1)})$ for the excess test error $\mathbb{E}[\mathcal{R}_M(\mathbf{v}_L)] - \sigma^2$. This matches the optimal rate for one-pass SGD (Lin et al., 2024) given the same compute budget, but requires only $N = \mathcal{O}(C^{b/(a+1)})$ number of i.i.d. samples in contrast to $N = \mathcal{O}(C^{a/(a+1)})$ for one-pass SGD.

Minimax optimal rate. When $a > b > 2$ and $M \gg N^{1/b}$, the improved rate $\widetilde{\mathcal{O}}(N^{(1-b)/b})$ achieved by multi-pass SGD matches the minimax optimal rate for a class of linear regression problems with similar spectral conditions (Pillaud-Vivien et al., 2018), up to sub-polynomial factors.

When $a < b < a + 1$, similarly, we have the following corollary from Theorem 3.1.

Corollary 3.4 (Scaling laws for multi-pass SGD when $a < b < a + 1$)**.** *Suppose the assumptions in Theorem 3.1 are in force and $\sigma^2 \eqsim 1$. When $a < b < a + 1$, for any $L_{\texttt{eff}} \lesssim N/\gamma$, we have*

$$\mathbb{E}[\mathcal{R}_M(\mathbf{v}_L)] = \sigma^2 + \Theta\bigg(\frac{1}{\min\{M, (L_{\texttt{eff}}\gamma)^{1/a}\}^{b-1}}\bigg) + \Theta\bigg(\frac{\min\{M, (L_{\texttt{eff}}\gamma)^{1/a}\}}{N}\bigg) + \mathbb{E}[\mathsf{Fluc}]$$

[2]Note that this choice of $L_{\texttt{eff}}$ (and therefore L) is optimal as it minimizes $\mathbb{E}[\mathcal{R}_M(\mathbf{v}_L)] - \sigma^2$ up to logarithmic factors for $L_{\texttt{eff}} \lesssim N^a$.

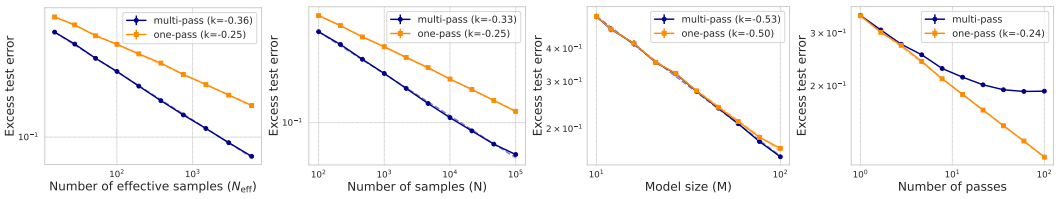
with probability at least $1 - e^{-\Omega(M)}$*, where the fluctuation error satisfies* $\mathbb{E}[\mathsf{Fluc}] \lesssim \gamma \log N \cdot \big(L_{\mathtt{eff}}\gamma\big)^{1/a-1}$*, and* $\mathbb{E}[\mathsf{Fluc}] \gtrsim \gamma\big(L_{\mathtt{eff}}\gamma\big)^{1/a-1}$ *when* $(L_{\mathtt{eff}}\gamma)^{1/a} \lesssim M$*.*

Therefore, in the data-constrained regime where $N \ll M^b$, we have $\mathbb{E}[\mathcal{R}_M(\mathbf{v}_L)] - \sigma^2 = \widetilde{\Theta}((L_{\mathtt{eff}}\gamma)^{(1-b)/a} + \gamma(L_{\mathtt{eff}}\gamma)^{1/a-1})$. Choosing $L_{\mathtt{eff}} \eqsim N$ and the optimal learning rate $\gamma \eqsim {L_{\mathtt{eff}}}^{a/b-1}$ that balances the excess test error of GD and the fluctuation error, we obtain a rate of $\widetilde{\Theta}(N^{(1-b)/b})$. This matches the bound for one-pass SGD in Lin et al. (2024) (up to logarithmic factors) when $a < b < a + 1$.

4 Experiments

We also perform simulations to validate our theoretical findings. Namely, we train M-dimensional sketched linear predictors (1) via one-pass SGD and multi-pass SGD following the setup in Section 2 and 3, and analyze how their excess test errors scale with the number of samples N and the model size M. In each simulation, we generate N i.i.d. samples $(\mathbf{x}_i, y_i)_{i=1}^N$ from a linear model $y_i = \langle \mathbf{x}_i, \mathbf{w}^* \rangle + \epsilon_i$, where $\mathbf{w}^* \in \mathbb{R}^d$ is an unknown parameter vector and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ are i.i.d. Gaussian noise. The covariates $\mathbf{x}_i$ are drawn from $\mathcal{N}(0, \mathbf{H})$, and the true parameter vector $\mathbf{w}^*$ is sampled from a Gaussian prior $\mathcal{N}(0, \mathbf{H}^{\mathbf{w}})$, where $\mathbf{H} := \mathrm{diag}\{1, 2^{-a}, \ldots, d^{-a}\}/\sum_{i=1}^d i^{-a}$ and $\mathbf{H}^{\mathbf{w}} := \mathrm{diag}\{1, 2^{a-b}, \ldots, d^{a-b}\}$ for some $a, b > 1$. We set the dimension d to be sufficiently large relative to M so that Assumption 1C and 1D are approximately satisfied. For simplicity, we implement multi-pass SGD by reusing samples sequentially without replacement in each epoch, rather than sampling i.i.d. from the empirical distribution. In all experiments, we set $d = 10000, \sigma^2 = 1$ and $(a, b) = (2, 1.5)$.

Figure 1(a) compares the excess test error of one-pass SGD and multi-pass SGD with the number of steps $L \eqsim N^{a/b-1}$. We observe that multi-pass SGD achieves better scaling in the sample size N compared to one-pass SGD when N is relatively small (i.e., $N \ll M^b$). Moreover, the fitted exponents are close to the theoretical predictions in Corollary 3.3 (i.e., $\frac{1-b}{a} = -0.25$ and $\frac{1-b}{b} = -0.33$). Similar results hold for the average of the iterates of constant-stepsize SGD, as shown in Figure 1(b). On the other hand, when $N \gg M^b$, Figure 1(c) shows that one-pass SGD and multi-pass SGD achieve the same scaling in the model size M with the exponent $k \approx 1 - b$, consistent with Corollary 3.3. In addition, Figure 1(d) illustrates that multi-pass SGD achieves the same excess test error as one-pass SGD on fresh data when the number of passes is below a certain threshold. Overall, the empirical observations align closely with our theoretical predictions.



(a) $\gamma = 0.1$ (b) $\gamma = 0.1$ (c) $\gamma = 0.5$ (d) $\gamma = 0.5, N = 1000$

Figure 1: Multi-pass SGD versus one-pass SGD. In (a)—(c), multi-pass SGD is ran for $L \eqsim N^{a/b}$ steps. (a), (b), (d): SGD with geometrically decaying stepsizes; (c): SGD with constant stepsizes. We use linear functions to fit the excess test error in log-log scale. The fitted exponents (k) are close to the theoretical predictions in Corollary 3.3. The errorbars denote the ± 1 standard deviation of the expected excess test error over 100 i.i.d. samples of $(\mathbf{S}, \mathbf{w}^*)$. Parameters: $\sigma^2 = 1$, $d = 10000$, $(a, b) = (2, 1.5)$. (a), (b), (d): $M = 1000$; (c): $N = 10^5$.

5 Proof Overview

We provide an overview of the proof of Theorem 3.1 in this section. A full proof can be found in Appendix A.2.1. Let $\mathbf{v}^* = (\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathbf{S}\mathbf{H}\mathbf{w}^*$ and adopt the shorthand notations

$$\mathbf{\Sigma} := \mathbf{S}\mathbf{H}\mathbf{S}^\top, \quad \widehat{\mathbf{\Sigma}} := \frac{\mathbf{S}\mathbf{X}^\top\mathbf{X}\mathbf{S}^\top}{N}.$$

It can be verified by some basic algebra that

$$\mathbb{E}\mathcal{R}_M(\mathbf{v}_L) = \mathbb{E}\Big[\big(\langle \mathbf{x}, \mathbf{S}^\top\mathbf{v}_L\rangle - y\big)^2\Big] \;\; = \underbrace{\sigma^2}_{\textsf{Irreducible}} + \mathbb{E}\underbrace{\|\mathbf{S}^\top\mathbf{v}^* - \mathbf{w}^*\|_{\mathbf{H}}^2}_{\textsf{Approx}} + \mathbb{E}\underbrace{\|\boldsymbol{\theta}_L - \mathbf{v}^*\|_{\mathbf{\Sigma}}^2}_{\textsf{Excess}} + \mathbb{E}\underbrace{\|\mathbf{v}_L - \boldsymbol{\theta}_L\|_{\mathbf{\Sigma}}^2}_{\textsf{Fluc}},$$

where the expectations on the R.H.S. are over $\mathbf{w}^*, (\mathbf{x}_i, y_i)_{i=1}^N$ and $(i_t)_{t=1}^L$, and we recall $\mathbf{v}_L$ in (multi-pass SGD) and $\boldsymbol{\theta}_L$ in (GD). From the above decomposition, we immediately have

1. $\textsf{Irreducible} = \mathcal{R}(\mathbf{w}^*) = \sigma^2$.
2. $\mathbb{E}_{\mathbf{w}^*}[\textsf{Approx}] = \mathbb{E}_{\mathbf{w}^*}\|\mathbf{S}^\top\mathbf{v}^* - \mathbf{w}^*\|_{\mathbf{H}}^2 \eqsim M^{1-b}$ with probability at least $1 - e^{-\Omega(M)}$ over $\mathbf{S}$ by Lemma C.5 in Lin et al. (2024) (see also Lemma E.7).

The excess risk of (GD) can be further decomposed into the sum of bias and variance, namely,

$$\mathbb{E}[\textsf{Excess}] = \textsf{Bias} + \sigma^2\textsf{Var},$$

where

$$\textsf{Bias} := \mathbb{E}\Big\|\prod_{t=1}^{L}\Big(\mathbf{I} - \gamma_t\widehat{\mathbf{\Sigma}}\Big)\mathbf{v}^*\Big\|_{\mathbf{\Sigma}}^2, \;\; \textsf{Var} := \mathbb{E}[\mathrm{tr}(\mathbf{X}\mathbf{S}^\top\mathbf{V}(\widehat{\mathbf{\Sigma}})\mathbf{\Sigma}\mathbf{V}(\widehat{\mathbf{\Sigma}})\mathbf{S}\mathbf{X}^\top)]$$

with $\mathbf{V}(\widehat{\mathbf{\Sigma}}) := \frac{1}{N}[\mathbf{I} - \prod_{t=1}^{L}(\mathbf{I} - \gamma_t\widehat{\mathbf{\Sigma}})]/\widehat{\mathbf{\Sigma}}$. The bounds on the bias (Bias) and variance (Var) then follow immediately from Lemma B.3 and C.2, respectively. Lastly, the bounds on the fluctuation error follow from Lemma D.5.

The main technical difficulty of proving Theorem 3.1 lies in bounding the bias, variance, and fluctuation error terms. For bias and variance upper bounds, due to the non-commutativity of the population covariance $\mathbf{\Sigma}$ and the empirical covariance $\widehat{\mathbf{\Sigma}}$, we apply a covariance replacement trick (Lemma E.1; see also Lemma 7 in Pillaud-Vivien et al. (2018)) to replace the population covariance with the empirical covariance in the expressions of bias and variance, as well as concentration properties of sub-Gaussian covariance to simplify their expressions. For the lower bounds, we show that a specific function of the empirical covariance commutes with the population covariance in expectation, and apply Von Neumann's trace inequality.

For the fluctuation error, we follow the standard practice as in Pillaud-Vivien et al. (2018) and Aguech et al. (2000) to express the difference between the multi-pass SGD and GD trajectories, $\mathbf{v}_t - \boldsymbol{\theta}_t$ as a stochastic process (Eq. (18)). We then bound the fluctuation error $\mathbb{E}[\|\mathbf{\Sigma}^{1/2}(\mathbf{v}_L - \boldsymbol{\theta}_L)\|^2]$ through controlling the accumulated error of the stochastic process using Lemma D.2 and D.3, which involves a novel leave-one-out argument to control the model parameters. Although several upper bounds on the fluctuation error have been established for infinite-dimensional linear models (Pillaud-Vivien et al., 2018; Zou et al., 2022), the interaction between the sketching matrix $\mathbf{S}$ and the samples $(\mathbf{x}_i, y_i)_{i=1}^N$ in our setup introduces additional technical challenges (Lin et al., 2024). Moreover, we derive a novel lower bound on the fluctuation error that matches the upper bound up to a logarithmic factor in certain regimes by carefully controlling the accumulated variance from random sampling (i.e. the accumulated variance induced by the random indices $(i_t)_{t=1}^L$).

6 Related Works

Empirical scaling laws. Scaling laws have been extensively studied in recent years as a way to understand and predict how model performance improves with increasing model size and data size (Hestness et al., 2017; Rosenfeld et al., 2019; Kaplan et al., 2020; Henighan et al., 2020; Hoffmann et al., 2022; Zhai et al., 2022; Muennighoff et al., 2023). The seminal work by Kaplan et al. (2020) introduced the concept of *neural scaling laws*, demonstrating empirically that the test

error of large transformer models decreases predictably following a power law with respect to the model size and data size. Subsequent works refined and extended these observations by proposing more accurate scaling formulas (Henighan et al., 2020; Hoffmann et al., 2022; Alabdulmohsin et al., 2022; Caballero et al., 2022; Muennighoff et al., 2023) and extending them to other settings (Kumar et al., 2024; Busbridge et al., 2025). In particular, Hoffmann et al. (2022) proposed the *Chinchilla scaling law*, which advocates scaling the model and data size proportionally as compute budget increases. Muennighoff et al. (2023) investigated the effect of data reuse and multiple training epochs, introducing an empirically refined scaling formula that accounts for the number of training epochs. They demonstrated that reused data can be approximately viewed as fresh data when the number of epochs is small.

Theoretical studies of scaling laws. Although scaling laws have been observed across diverse settings, their theoretical understanding remains relatively limited. A number of recent works have attempted to formalize and explain the observed scaling behaviors in simplified settings (Sharma and Kaplan, 2020; Bahri et al., 2021; Maloney et al., 2022; Hutter, 2021; Michaud et al., 2024; Bordelon et al., 2024a; Atanasov et al., 2024; Dohmatob et al., 2024; Paquette et al., 2024; Lin et al., 2024; Bordelon et al., 2024b; Ren et al., 2025). For example, Bahri et al. (2021) considered a linear teacher-student model with a power-law spectrum and showed that the test error of the ordinary least squares estimator scales following a power law in N (or M) when the other parameter goes to infinity. Bordelon et al. (2024a) analyzed the test error of the solution found by gradient flow in a linear random feature model and established power-law scaling in one of N, M and T (training time) while the other two parameters go to infinity. The results in these works are derived based on statistical physics heuristics and characterize scaling in only one variable in the asymptotic regime. More recently, Lin et al. (2024) analyzed the test error of the last iterate of one-pass SGD in a sketched linear model and showed that the test error scales as $\Theta(\sigma^2 + M^{1-b} + N^{(1-b)/a})$ under the source condition (Assumption 1). This is the first work to establish a finite-sample joint scaling law (in M and N) for linear models that aligns with empirical observations (Kaplan et al., 2020; Hoffmann et al., 2022). Similarly, Ren et al. (2025) analyzed the complexity of one-pass SGD for learning two-layer neural networks in a teacher-student setup, and derived joint scaling laws for the test error under power-law assumptions on the teacher network. While previous works study the scaling behavior of the one-pass (online) SGD solutions, our work complements them by analyzing the effect of data reuse (i.e., multi-pass SGD) in data-constrained regimes.

Risk bounds for SGD. The generalization behavior of stochastic gradient descent (SGD), particularly in linear regression, has been extensively studied across both classical and high-dimensional regimes (Polyak and Juditsky, 1992; Défossez and Bach, 2015; Dieuleveut et al., 2017; Jain et al., 2018, 2017; Pillaud-Vivien et al., 2018; Ge et al., 2019; Dieuleveut and Bach, 2015; Berthier et al., 2020; Zou et al., 2023, 2021, 2022; Wu et al., 2022b,c; Varre et al., 2021). For one-pass SGD, several works have developed tight test error bounds in overparameterized linear models (Zou et al., 2023; Wu et al., 2022a,c). For multi-pass SGD, early works (Lin and Rosasco, 2017; Pillaud-Vivien et al., 2018; Mücke et al., 2019; Zou et al., 2022) have established test error bounds for the average of its iterates in linear regression. Compared with prior works, our main technical contribution is to precisely control the effect of random sketching and to refine the characterization of fluctuation error (see Fluc in Eq. 3) in the multi-pass setting. Under comparable regimes where the approximation error is zero, our test error bounds match those derived in Pillaud-Vivien et al. (2018), which are minimax optimal for a specific class of linear regression problems in certain cases.

7 Conclusion

In this work, we provide a theoretical analysis of multi-pass stochastic gradient descent (multi-pass SGD) in a sketched linear regression problem and establish refined scaling laws that characterize how the test error scales with the model size M, sample size N, and number of optimization steps L. Our results show that, under suitable power-law conditions on the true parameter and data distribution, data reuse via multi-pass SGD can improve model performance when the number of samples is limited. This offers a theoretical explanation for the empirical benefits of multiple passes in modern large-scale training.

Our analysis has several limitations. One limitation is the assumption that the eigenvectors of the prior and data covariance are aligned (implied by Assumption 1D). While this assumption cannot be

fully removed without affecting the error rate, it would be interesting to investigate what alternative rates are achieved when the eigenvectors are not aligned. Another limitation is that our lower bound results require Gaussian design of the covariates (i.e., Assumption 1A); a next step is to extend them to non-Gaussian design.

Beyond the limitations, many other directions remain open for future research. First, our analysis focuses on multi-pass SGD with batch size one; it would be worthwhile to understand how the test error scales with the batch size and to develop corresponding batch size scaling laws (see Jain et al., 2017). Another important direction is to study how data reuse interacts with other optimization algorithms, such as SGD with momentum or ℓ_2-regularization and Adam. In addition, it is valuable to extend our analysis to non-linear settings and classification problems, such as logistic regression, kernel methods, and neural networks. Notably, modern large language model pretraining is based on minimizing the cross-entropy loss for next-word prediction. Understanding the scaling behavior in logistic regression—the simplest classification model—thus represents an important step toward unraveling the mysteries of LLM scaling.

Acknowledgements

We gratefully acknowledge the NSF's support of FODSI through grant DMS-2023505 and of the NSF and the Simons Foundation for the Collaboration on the Theoretical Foundations of Deep Learning through awards DMS-2031883 and #814639 and of the ONR through MURI award N000142112431.

Bibliography

Rafik Aguech, Eric Moulines, and Pierre Priouret. On a perturbation approach for the analysis of stochastic tracking algorithms. *SIAM Journal on Control and Optimization*, 39(3):872–899, 2000.

Ibrahim M Alabdulmohsin, Behnam Neyshabur, and Xiaohua Zhai. Revisiting neural scaling laws in language and vision. *Advances in Neural Information Processing Systems*, 35:22300–22312, 2022.

Alexander B Atanasov, Jacob A Zavatone-Veth, and Cengiz Pehlevan. Scaling and renormalization in high-dimensional regression. *arXiv preprint arXiv:2405.00592*, 2024.

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *arXiv preprint arXiv:2102.06701*, 2021.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 2020.

Raphaël Berthier, Francis Bach, and Pierre Gaillard. Tight nonparametric convergence rates for stochastic gradient descent under the noiseless linear model. *Advances in Neural Information Processing Systems*, 33:2576–2586, 2020.

Tamay Besiroglu, Ege Erdil, Matthew Barnett, and Josh You. Chinchilla scaling: A replication attempt. *arXiv preprint arXiv:2404.10102*, 2024.

Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. A dynamical model of neural scaling laws. *arXiv preprint arXiv:2402.01092*, 2024a.

Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. How feature learning can improve neural scaling laws. *arXiv preprint arXiv:2409.17858*, 2024b.

Dan Busbridge, Amitis Shidani, Floris Weers, Jason Ramapuram, Etai Littwin, and Russ Webb. Distillation scaling laws. *arXiv preprint arXiv:2502.08606*, 2025.

Ethan Caballero, Kshitij Gupta, Irina Rish, and David Krueger. Broken neural scaling laws. *arXiv preprint arXiv:2210.14891*, 2022.

Alexandre Défossez and Francis Bach. Averaged least-mean-squares: Bias-variance trade-offs and optimal sampling distributions. In *Artificial Intelligence and Statistics*, pages 205–213, 2015.

Aymeric Dieuleveut and Francis R. Bach. Non-parametric stochastic approximation with large step sizes. *The Annals of Statistics*, 2015.

Aymeric Dieuleveut, Nicolas Flammarion, and Francis Bach. Harder, better, faster, stronger convergence rates for least-squares regression. *The Journal of Machine Learning Research*, 18(1): 3520–3570, 2017.

Elvis Dohmatob, Yunzhen Feng, Pu Yang, Francois Charton, and Julia Kempe. A tale of tails: Model collapse as a change of scaling laws. *arXiv preprint arXiv:2402.07043*, 2024.

Rong Ge, Sham M Kakade, Rahul Kidambi, and Praneeth Netrapalli. The step decay schedule: A near optimal, geometrically decaying learning rate procedure for least squares. *Advances in neural information processing systems*, 32, 2019.

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, et al. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Daniel Hsu, Sham Kakade, and Tong Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17(none):1 – 6, 2012. doi: 10.1214/ECP.v17-2079. URL `https://doi.org/10.1214/ECP.v17-2079`.

Marcus Hutter. Learning curve theory. *arXiv preprint arXiv:2102.04074*, 2021.

Prateek Jain, Praneeth Netrapalli, Sham M Kakade, Rahul Kidambi, and Aaron Sidford. Parallelizing stochastic gradient descent for least squares regression: mini-batching, averaging, and model misspecification. *The Journal of Machine Learning Research*, 18(1):8258–8299, 2017.

Prateek Jain, Sham M. Kakade, Rahul Kidambi, Praneeth Netrapalli, Venkata Krishna Pillutla, and Aaron Sidford. A Markov Chain Theory Approach to Characterizing the Minimax Optimality of Stochastic Gradient Descent (for Least Squares). In *37th IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS 2017)*, 2018.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Vladimir Koltchinskii and Karim Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, pages 110–133, 2017.

Tanishq Kumar, Zachary Ankner, Benjamin F Spector, Blake Bordelon, Niklas Muennighoff, Mansheej Paul, Cengiz Pehlevan, Christopher Ré, and Aditi Raghunathan. Scaling laws for precision. *arXiv preprint arXiv:2411.04330*, 2024.

Junhong Lin and Lorenzo Rosasco. Optimal rates for multi-pass stochastic gradient methods. *The Journal of Machine Learning Research*, 18(1):3375–3421, 2017.

Licong Lin, Jingfeng Wu, Sham M Kakade, Peter L Bartlett, and Jason D Lee. Scaling laws in linear regression: Compute, parameters, and data. *arXiv preprint arXiv:2406.08466*, 2024.

Alexander Maloney, Daniel A Roberts, and James Sully. A solvable model of neural scaling laws. *arXiv preprint arXiv:2210.16859*, 2022.

Eric Michaud, Ziming Liu, Uzay Girit, and Max Tegmark. The quantization model of neural scaling. *Advances in Neural Information Processing Systems*, 36, 2024.

Nicole Mücke, Gergely Neu, and Lorenzo Rosasco. Beating sgd saturation with tail-averaging and minibatching. *Advances in Neural Information Processing Systems*, 32, 2019.

Niklas Muennighoff, Alexander M Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. Scaling data-constrained language models. *arXiv preprint arXiv:2305.16264*, 2023.

Elliot Paquette, Courtney Paquette, Lechao Xiao, and Jeffrey Pennington. 4+ 3 phases of compute-optimal neural scaling laws. *arXiv preprint arXiv:2405.15074*, 2024.

Steven T Piantadosi. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, 21:1112–1130, 2014.

Loucas Pillaud-Vivien, Alessandro Rudi, and Francis Bach. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. *Advances in Neural Information Processing Systems*, 31, 2018.

Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

Yunwei Ren, Eshaan Nichani, Denny Wu, and Jason D Lee. Emergence and scaling laws in sgd learning of shallow neural networks. *arXiv preprint arXiv:2504.19983*, 2025.

Jonathan S Rosenfeld, Amir Rosenfeld, Yonatan Belinkov, and Nir Shavit. A constructive prediction of the generalization error across scales. *arXiv preprint arXiv:1909.12673*, 2019.

Utkarsh Sharma and Jared Kaplan. A neural scaling law from the dimension of the data manifold. *arXiv preprint arXiv:2004.10802*, 2020.

Aditya Vardhan Varre, Loucas Pillaud-Vivien, and Nicolas Flammarion. Last iterate convergence of SGD for least-squares in the interpolation regime. In *Advances in Neural Information Processing Systems*, 2021.

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

Jingfeng Wu, Difan Zou, Vladimir Braverman, Quanquan Gu, and Sham Kakade. Last iterate risk bounds of sgd with decaying stepsize for overparameterized linear regression. In *International Conference on Machine Learning*, pages 24280–24314. PMLR, 2022a.

Jingfeng Wu, Difan Zou, Vladimir Braverman, Quanquan Gu, and Sham M. Kakade. Last iterate risk bounds of sgd with decaying stepsize for overparameterized linear regression. *The 39th International Conference on Machine Learning*, 2022b.

Jingfeng Wu, Difan Zou, Vladimir Braverman, Quanquan Gu, and Sham M. Kakade. The power and limitation of pretraining-finetuning for linear regression under covariate shift. *The 36th Conference on Neural Information Processing Systems*, 2022c.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12104–12113, 2022.

Nikita Zhivotovskiy. Dimension-free bounds for sums of independent matrices and simple tensors via the variational principle. *Electronic Journal of Probability*, 29:1–28, 2024.

Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, Dean P Foster, and Sham Kakade. The benefits of implicit regularization from sgd in least squares problems. *Advances in Neural Information Processing Systems*, 34:5456–5468, 2021.

Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, and Sham Kakade. Risk bounds of multi-pass sgd for least squares in the interpolation regime. *Advances in Neural Information Processing Systems*, 35:12909–12920, 2022.

Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, and Sham M Kakade. Benign overfitting of constant-stepsize sgd for linear regression. *Journal of Machine Learning Research*, 24(326):1–58, 2023.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: A summary of our results and contributions is provided in the introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: we have discussed the limitations of our work in the paper.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: it can be found from our theorem statements and the proofs.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: we have discussed all experimental details for reproducing the simulation results in the paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: we do not release code and data for this paper at this time.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: we have specified all the training and test details in the paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: we have reported error bars for all experiments in the paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: the simulations in this paper do not involve any large language models and can be reproduced on a personal computer.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: the authors have reviewed the Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: there is no societal impact of the work performed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: not applicable

    Guidelines:

    - The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: the paper does not use existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: the paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: this paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: this paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: the core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

Appendix

Table of Contents

A Preliminary

A.1 Comments and additional notations

Comments on Assumption 1D. Throughout the appendix (except for Appendix A.3), we assume without loss of generality that the covariance matrix $\mathbf{H}$ is diagonal, with diagonal entries given by the eigenvalues $(\lambda_i)_{i\geqslant 1}$ in non-increasing order. This reduction is justified by the rotational invariance of the Gaussian sketching matrix $\mathbf{S}$. Under this diagonalization, Assumption 1D can be restated more explicitly as follows:

Assumption 2 (Source condition)**.** *Suppose $\mathbf{H} = (h_{ij})_{i,j\geqslant 1}$ is a diagonal matrix with non-increasing diagonal entries. Assume that the true parameter $\mathbf{w}^*$ satisfies:*

$$\textit{for all } i \neq j, \quad \mathbb{E}[\mathbf{w}_i^* \mathbf{w}_j^*] = 0; \qquad \textit{and for all } i > 0, \quad \mathbb{E}[\lambda_i \mathbf{w}_i^{*2}] \eqsim i^{-b}, \quad \textit{for some } b > 1.$$

Given that $\mathbf{H}$ is diagonal, we adopt the following notation. For integers $0 \leqslant k^* \leqslant k^\dagger$ (allowing $k^\dagger = \infty$), define

$$\mathbf{H}_{k^*:k^\dagger} := \mathrm{diag}\{\lambda_{k^*+1}, \ldots, \lambda_{k^\dagger}\} \in \mathbb{R}^{(k^\dagger - k^*)\times(k^\dagger - k^*)}.$$

For example,

$$\mathbf{H}_{0:k} = \mathrm{diag}\{\lambda_1, \ldots, \lambda_k\}, \quad \mathbf{H}_{k:\infty} = \mathrm{diag}\{\lambda_{k+1}, \lambda_{k+2}, \ldots\}.$$

Similarly, for any vector $\mathbf{w} \in \mathbb{H}$, define

$$\mathbf{w}_{k^*:k^\dagger} := (\mathbf{w}_{k^*+1}, \ldots, \mathbf{w}_{k^\dagger})^\top \in \mathbb{R}^{k^\dagger - k^*}.$$

In addition, we define $\mathbf{S}_{k^*:k^\dagger}$ to be the submatrix of the sketching matrix $\mathbf{S}$ consisting of the k^*+1-th through $k^\dagger$-th columns.

A.1.1 Assumptions on the stepsize

In the proofs of the general upper and lower bounds on the bias, variance, and fluctuation error, we will require that the stepsize γ_0, γ satisfy certain conditions, which are summarized in the following assumption.

Assumption 3 (Stepsize conditions)**.** *Under the notations in Theorem 3.1 and its proof, with probability at least* $1 - \exp(-\Omega(M))$ *over the randomness of* $\mathbf{S}$*, we have*

1. $\gamma \leqslant \min\{c/\log N, c/[\mathrm{tr}(\mathbf{\Sigma})]\}$;
2. $\mathrm{tr}(\mathbf{\Sigma}^2) \lesssim 1$;
3. $\sum_{i=1}^{M} \frac{\mu_i(\mathbf{\Sigma})}{\mu_i(\mathbf{\Sigma}) + 1/(L_{\mathtt{eff}}\gamma)} \leqslant N/4$;
4. *the initial stepsize* $\gamma_0 = \min\{1/[4\max_i \|\mathbf{S}\mathbf{x}_i\|_2^2], \gamma\}$ *satisfies* $\mathbb{P}(\gamma_0 < \gamma/t) \leqslant N^{-ct}$ *for all* $t \geqslant 1$.

We will show that Assumption 3 holds when the conditions in Theorem 3.1 are satisfied.

A.2 Proof of Theorem 3.1 and the corollaries

A.2.1 Proof of Theorem 3.1

Proof of Theorem 3.1. Let $\mathbf{v}^* = (\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathbf{S}\mathbf{H}\mathbf{w}^*$ and adopt the shorthand

$$\mathbf{\Sigma} := \mathbf{S}\mathbf{H}\mathbf{S}^\top, \quad \widehat{\mathbf{\Sigma}} := \frac{\mathbf{S}\mathbf{X}^\top\mathbf{X}\mathbf{S}^\top}{N}.$$

Also, let $\mathcal{D} := (\mathbf{x}_i, y_i)_{i=1}^N$ denote the set of training samples. Then we have the decomposition

$$\begin{aligned}
\mathbb{E}\mathcal{R}_M(\mathbf{v}_L) &= \mathbb{E}\Big[\big(\langle \mathbf{x}, \mathbf{S}^\top\mathbf{v}_L\rangle - y\big)^2\Big] = \mathbb{E}\Big[\big(\langle \mathbf{x}, \mathbf{S}^\top\mathbf{v}_L - \mathbf{w}^*\rangle - \epsilon\big)^2\Big] = \sigma^2 + \mathbb{E}\big[\langle \mathbf{x}, \mathbf{S}^\top\mathbf{v}_L - \mathbf{w}^*\rangle^2\big] \\
&= \sigma^2 + \mathbb{E}\big[\langle \mathbf{x}, \mathbf{S}^\top(\mathbf{v}_L - \mathbf{v}^*) + \mathbf{S}^\top\mathbf{v}^* - \mathbf{w}^*\rangle^2\big] \\
&\overset{(i)}{=} \sigma^2 + \mathbb{E}\big[\langle \mathbf{x}, \mathbf{S}^\top\mathbf{v}^* - \mathbf{w}^*\rangle^2\big] + \mathbb{E}\big[\langle \mathbf{x}, \mathbf{S}^\top(\mathbf{v}_L - \mathbf{v}^*)\rangle^2\big] \\
&\overset{(ii)}{=} \sigma^2 + \mathbb{E}\big[\langle \mathbf{x}, \mathbf{S}^\top\mathbf{v}^* - \mathbf{w}^*\rangle^2\big] + \mathbb{E}\big[\langle \mathbf{x}, \mathbf{S}^\top(\boldsymbol{\theta}_L - \mathbf{v}^*)\rangle^2\big] + \mathbb{E}\big[\langle \mathbf{x}, \mathbf{S}^\top(\mathbf{v}_L - \boldsymbol{\theta}_L)\rangle^2\big] \\
&= \underbrace{\sigma^2}_{\mathsf{Irreducible}} + \mathbb{E}\underbrace{\|\mathbf{S}^\top\mathbf{v}^* - \mathbf{w}^*\|_{\mathbf{H}}^2}_{\mathsf{Approx}} + \mathbb{E}\underbrace{\|\boldsymbol{\theta}_L - \mathbf{v}^*\|_{\mathbf{\Sigma}}^2}_{\mathsf{Excess}} + \mathbb{E}\underbrace{\|\mathbf{v}_L - \boldsymbol{\theta}_L\|_{\mathbf{\Sigma}}^2}_{\mathsf{Fluc}},
\end{aligned}$$

where step (i) uses the fact that $\mathbb{E}[\mathbf{S}\mathbf{x}\mathbf{x}^\top(\mathbf{S}^\top\mathbf{v}^* - \mathbf{w}^*)] = \mathbb{E}[\mathbf{S}\mathbf{H}\mathbf{S}^\top\mathbf{v}^* - \mathbf{S}\mathbf{H}\mathbf{w}^*] = 0$, and step (ii) uses the fact that $\mathbb{E}[\mathbf{v}_L|\mathbf{S}, \mathbf{w}^*, \mathcal{D}] = \boldsymbol{\theta}_L$.

Irreducible error. From the above decomposition, we have $\mathsf{Irreducible} = \mathcal{R}(\mathbf{w}^*) = \sigma^2$.

Approximation error. We have from Lemma C.5 in Lin et al. (2024) that $\mathbb{E}_{\mathbf{w}^*}\mathsf{Approx} = \mathbb{E}_{\mathbf{w}^*}\|\mathbf{S}^\top\mathbf{v}^* - \mathbf{w}^*\|_{\mathbf{H}}^2 \eqsim M^{1-b}$ with probability at least $1 - e^{-\Omega(M)}$ over $\mathbf{S}$.

Excess risk of (GD)**.** Let $\tilde{\epsilon}_i = y_i - \mathbf{x}_i^\top\mathbf{S}^\top\mathbf{v}^*$ for $i \in [N]$ and write $\tilde{\boldsymbol{\epsilon}} = (\tilde{\epsilon}_1, \ldots, \tilde{\epsilon}_N)^\top$. It can be verified that, conditioned on $(\mathbf{S}, \mathbf{w}^*)$, $\mathbb{E}[\tilde{\epsilon}_i] = 0$ and $\tilde{\epsilon}_i$ is independent of $\mathbf{S}\mathbf{x}_i$. Moreover,

$$\begin{aligned}
\sigma^2 \leqslant \tilde{\sigma}^2 := \mathbb{E}[\tilde{\epsilon}_i^2] &= \sigma^2 + \mathbb{E}_{\mathbf{w}^*}\|\mathbf{w}^* - \mathbf{S}^\top\mathbf{v}^*\|_{\mathbf{H}}^2 \\
&= \sigma^2 + \mathbb{E}_{\mathbf{w}^*}[\mathbf{w}^{*\top}\mathbf{H}^{1/2}(\mathbf{I} - \mathbf{H}^{1/2}\mathbf{S}^\top(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathbf{S}\mathbf{H}^{1/2})\mathbf{H}^{1/2}\mathbf{w}^*] \\
&\leqslant \sigma^2 + \mathbb{E}_{\mathbf{w}^*}\|\mathbf{w}^*\|_{\mathbf{H}}^2 \lesssim \sigma^2.
\end{aligned}$$

Note that by definition of (GD), we have

$$\begin{aligned}
\boldsymbol{\theta}_t - \mathbf{v}^* &= \boldsymbol{\theta}_t - \mathbf{v}^* - \frac{\gamma_t}{N}\mathbf{S}\mathbf{X}^\top(\mathbf{X}\mathbf{S}^\top\boldsymbol{\theta}_{t-1} - \mathbf{y}) = \boldsymbol{\theta}_{t-1} - \mathbf{v}^* - \frac{\gamma_t}{N}\mathbf{S}\mathbf{X}^\top(\mathbf{X}\mathbf{S}^\top(\boldsymbol{\theta}_{t-1} - \mathbf{v}^*) - \tilde{\boldsymbol{\epsilon}}) \\
&= \Big(\mathbf{I} - \gamma_t\widehat{\mathbf{\Sigma}}\Big)(\boldsymbol{\theta}_{t-1} - \mathbf{v}^*) + \frac{\gamma_t}{N} \cdot \mathbf{S}\mathbf{X}^\top\tilde{\boldsymbol{\epsilon}},
\end{aligned}$$

and therefore

$$\boldsymbol{\theta}_L - \mathbf{v}^* = \prod_{t=1}^{L} \Big(\mathbf{I} - \gamma_t \widehat{\boldsymbol{\Sigma}}\Big)(\boldsymbol{\theta}_0 - \mathbf{v}^*) + \mathbf{V}(\widehat{\boldsymbol{\Sigma}})\mathbf{S}\mathbf{X}^\top \tilde{\boldsymbol{\epsilon}}, \tag{4}$$

where

$$\mathbf{V}(\widehat{\boldsymbol{\Sigma}}) := \frac{1}{N}\sum_{t=1}^{L} \gamma_t \cdot \prod_{i=t+1}^{L} (\mathbf{I} - \gamma_i \widehat{\boldsymbol{\Sigma}}) = \frac{\mathbf{I} - \prod_{t=1}^{L}(\mathbf{I} - \gamma_t \widehat{\boldsymbol{\Sigma}})}{N\widehat{\boldsymbol{\Sigma}}}.$$

As a result, the excess risk of (GD) satifies

$$\begin{aligned}\mathbb{E}[\mathsf{Excess}] &= \mathbb{E}\|\boldsymbol{\theta}_L - \mathbf{v}^*\|_{\boldsymbol{\Sigma}}^2 \overset{(iii)}{=} \mathbb{E}\Big\| \prod_{t=1}^{L} \Big(\mathbf{I} - \gamma_t \widehat{\boldsymbol{\Sigma}}\Big)\mathbf{v}^*\Big\|_{\boldsymbol{\Sigma}}^2 + \mathbb{E}\|\mathbf{V}(\widehat{\boldsymbol{\Sigma}})\mathbf{S}\mathbf{X}^\top\tilde{\boldsymbol{\epsilon}}\|_{\boldsymbol{\Sigma}}^2 \\ &\eqsim \mathsf{Bias} + \sigma^2 \mathsf{Var},\end{aligned}$$

where $\mathsf{Bias} := \mathbb{E}_{\mathbf{w}*}[\mathsf{Bias}(\mathbf{w}^*)]$ and

$$\mathsf{Bias}(\mathbf{w}^*) := \mathbb{E}_{\mathbf{X}}\Big\| \prod_{t=1}^{L} \Big(\mathbf{I} - \gamma_t \widehat{\boldsymbol{\Sigma}}\Big)\mathbf{v}^*\Big\|_{\boldsymbol{\Sigma}}^2, \qquad \mathsf{Var} := \mathbb{E}[\mathrm{tr}(\mathbf{X}\mathbf{S}^\top \mathbf{V}(\widehat{\boldsymbol{\Sigma}})\boldsymbol{\Sigma}\mathbf{V}(\widehat{\boldsymbol{\Sigma}})\mathbf{S}\mathbf{X}^\top)],$$

and step (iii) follows from the fact that $\mathbf{S}\mathbf{X}^\top$ is independent of $\tilde{\boldsymbol{\epsilon}}$ conditioned on $\mathbf{S}$. The bounds on the bias and variance follow immediately from Lemma B.3 and C.2.

Fluctuation error. It follows from Lemma D.5 and the assumption $\gamma \leqslant c/\log N$ that

$$\mathbb{E}[\mathsf{Fluc}] \lesssim \gamma \log N \cdot [(L_{\mathtt{eff}}\gamma)^{1/a-1} + \frac{(L_{\mathtt{eff}}\gamma)^{1/a}}{N}]$$

with probability at least $1 - e^{-\Omega(M)}$ over the randomness of $\mathbf{S}$. The lower bound on $\mathbb{E}[\mathsf{Fluc}]$ also follows from Lemma D.5.

□

A.2.2 Proof of Corollary 3.2

The proof follows immediately by combining parts 1–3 of Theorem 3.1, although we make a different assumption on the initial stepsize γ_0. In Theorem 3.1, we assume $\gamma_0 = \min\{\gamma, 1/[4\max_i \|\mathbf{S}\mathbf{x}_i\|_2^2]\}$ for some $\gamma \lesssim 1/\log N$, while in Corollary 3.2, we assume $\gamma_0 = \min\{\gamma, 1/[4\,\mathrm{tr}(\mathbf{S}\mathbf{X}^\top\mathbf{X}\mathbf{S}^\top/N)]\}$ for some $\gamma \lesssim 1$. This modification is valid because Lemmas B.1, B.2, and C.1, used in the proof of parts 1–3 of Theorem 3.1, continue to hold under the alternative choice of stepsize.

Specifically, their proofs mainly rely on three properties: (1) $\mathbf{I} - \gamma_t \mathbf{S}\mathbf{X}^\top\mathbf{X}\mathbf{S}^\top/N \succeq \mathbf{0}$, (2) $\mathbb{P}(\gamma_0 < \gamma/t) \leqslant N^{-ct}$ for all $t \geqslant 1$ and (3) claim (15a) holds. Under the choice $\gamma_0 = \min\{\gamma, 1/[4\,\mathrm{tr}(\mathbf{S}\mathbf{X}^\top\mathbf{X}\mathbf{S}^\top/N)]\}$, the first two properties are satisfied by definition and by the Hanson–Wright inequality (see, e.g., exercise 2.17 in Wainwright (2019)). The third property follows from a similar symmetry property for $\gamma_0(\Gamma) := \min\{1/[4\,\mathrm{tr}(\Gamma\Gamma^\top/N)], \gamma\}$ as used in the proof of claim (15a).

A.2.3 Proof of Corollary 3.3 and 3.4

These two corollaries follow immediately from combining parts 1–4 of Theorem 3.1 and some basic algebra.

A.3 Relaxation of Assumption 1

In this section, we show that some conditions in Assumption 1 can be further relaxed. Concretely, we have

(a). The exact alignment of the eigenvectors of the prior and data covariance in Assumption 1D is not necessary. All results in Section 3 remain valid if Assumption 1D is replaced by

Assumption 1D' (Approximate source condition)**.** *Let* $(\lambda_i, \mathbf{v}_i)_{i>0}$ *be the eigenvalues and eigenvectors of* $\mathbf{H}$ *and let* $\mathbf{H}^{\mathbf{w}} = \mathbb{E}[\mathbf{w}^*\mathbf{w}^{*\top}]$*. Assume* $c\widetilde{\mathbf{H}}^{\mathbf{w}} \preceq \mathbf{H}^{\mathbf{w}} \preceq c'\widetilde{\mathbf{H}}^{\mathbf{w}}$ *for some absolute constants* $c' \geqslant c > 0$ *and* $\widetilde{\mathbf{H}}^{\mathbf{w}} \succeq 0$ *such that*

$$\text{for } i \neq j,\ \ \mathbf{v}_i^\top \widetilde{\mathbf{H}}^{\mathbf{w}} \mathbf{v}_j = 0;\ \ \text{and for } i > 0,\ \ \lambda_i \mathbf{v}_i^\top \widetilde{\mathbf{H}}^{\mathbf{w}} \mathbf{v}_i \eqsim i^{-b},\ \ \text{for some } b > 1.$$

(b). To establish the upper bounds for Bias, Var, Approx in Theorem 3.1 and the upper bounds in Corollary 3.2—3.4, Assumption 1A can be relaxed to

Assumption 1A' (sub-Gaussian design)**.** $\mathbf{x} = \mathbf{H}^{1/2}\widetilde{\mathbf{x}}$*, where* $\mathbb{E}[\widetilde{\mathbf{x}}\widetilde{\mathbf{x}}^\top] = \mathbf{I}$*, and the vector* $\widetilde{\mathbf{x}}$ *is zero-mean and 1-sub-Gaussian, i.e.,* $\mathbb{E}[\widetilde{\mathbf{x}}] = \mathbf{0}$ *and* $\mathbb{E}[e^{\lambda\langle \mathbf{v}, \widetilde{\mathbf{x}}\rangle}] \leqslant e^{\lambda^2/2}$ *for any unit vector* $\mathbf{v}$ *and all* $\lambda \in \mathbb{R}$.

We provide some justification of the two relaxations below.

Justification of (a). By checking the proof of Theorem 3.1 and its corollaries, it can be seen that Assumption 1D is used to (1) give matching upper and lower bounds on $\mathbb{E}_{\mathbf{w}^*}[\|\mathbf{w}^*_{0:k}\|_2^2], \mathbb{E}_{\mathbf{w}^*}[\|\mathbf{w}^*_{k:\infty}\|^2_{\mathbf{H}_{k:\infty}}], \mu_i(\mathbf{S}\mathbf{H}\mathbf{H}^{\mathbf{w}}\mathbf{H}\mathbf{S}^\top)$ for any $k \geqslant 0$ and $i \in [M]$ when controlling the approximation and bias error (see Lemma C.5 in Lin et al. (2024) and Lemma B.3); (2) give matching upper and lower bounds on $\mathbb{E}[\|\mathbf{w}^*\|^2_{\mathbf{H}}]$ when controlling the fluctuation error (see Lemma D.5). Under the alternative Assumption 1D', it is readily verified that the same bounds on these quantities can be established up to constant factors. Concretely, suppose there exists some parameter $\widetilde{\mathbf{w}}^*$ with prior $\mathbb{E}[\widetilde{\mathbf{w}}^*\widetilde{\mathbf{w}}^{*\top}] = \widetilde{\mathbf{H}}^{\mathbf{w}}$. Then $\widetilde{\mathbf{w}}^*$ satisfies Assumption 1D and

$$\begin{aligned}
\mathbb{E}_{\mathbf{w}^*}[\|\mathbf{w}^*_{0:k}\|_2^2] = \mathrm{tr}(\mathbf{H}^{\mathbf{w}}_{0:k}) \eqsim \mathrm{tr}(\widetilde{\mathbf{H}}^{\mathbf{w}}_{0:k}) = \mathbb{E}_{\widetilde{\mathbf{w}}^*}[\|\widetilde{\mathbf{w}}^*_{0:k}\|_2^2],\\
\mathbb{E}_{\mathbf{w}^*}[\|\mathbf{w}^*_{k:\infty}\|^2_{\mathbf{H}_{k:\infty}}] = \mathrm{tr}(\mathbf{H}^{1/2}_{k:\infty}\mathbf{H}^{\mathbf{w}}_{k:\infty}\mathbf{H}^{1/2}_{k:\infty}) \eqsim \mathrm{tr}(\mathbf{H}^{1/2}_{k:\infty}\widetilde{\mathbf{H}}^{\mathbf{w}}_{k:\infty}\mathbf{H}^{1/2}_{k:\infty}) = \mathbb{E}_{\widetilde{\mathbf{w}}^*}[\|\widetilde{\mathbf{w}}^*_{k:\infty}\|^2_{\mathbf{H}_{k:\infty}}],\\
\mu_i(\mathbf{S}\mathbf{H}\mathbf{H}^{\mathbf{w}}\mathbf{H}\mathbf{S}^\top) \overset{(i)}{\eqsim} \mu_i(\mathbf{S}\mathbf{H}\widetilde{\mathbf{H}}^{\mathbf{w}}\mathbf{H}\mathbf{S}^\top),\\
\mathbb{E}[\|\mathbf{w}^*\|^2_{\mathbf{H}}] = \mathrm{tr}(\mathbf{H}^{1/2}\mathbf{H}^{\mathbf{w}}\mathbf{H}^{1/2}) \eqsim \mathrm{tr}(\mathbf{H}^{1/2}\widetilde{\mathbf{H}}^{\mathbf{w}}\mathbf{H}^{1/2}) = \mathbb{E}_{\widetilde{\mathbf{w}}^*}[\|\widetilde{\mathbf{w}}^*\|^2_{\mathbf{H}}],
\end{aligned}$$

where step (i) follows from the fact that $\mu_i(\mathbf{A}) \leqslant \mu_i(\mathbf{B})$ for all i and any $\mathbf{0} \preceq \mathbf{A} \preceq \mathbf{B}$. Therefore, the proof of Theorem 3.1 and its corollaries goes through under the alternative Assumption 1D'.

Justification of (b). In short, for the upper bounds, the relaxation can be made since the Gaussian assumption is mainly used to establish certain concentration bounds (e.g., Bernstein's inequality), which also hold for sub-Gaussian vectors. More specifically, the Gaussian design in Assumption 1A' is used in our proof mainly in three ways: (1) to establish concentration bounds on the sample covariance (e.g., Eq. 10); (2) to allow the use of technical lemmas in Appendix E (e.g., Lemma E.3 and E.4); (3) to control the norm of sketched samples (e.g., to control $B_{\boldsymbol{\nu}}$ in Eq. 20).

Correspondingly, when $\mathbf{x}$ satisfies the alternative Assumption 1A', we can show that (1) the same concentration bounds hold on the sub-Gaussian sample covariance by e.g., Theorem 6.5 in Wainwright (2019); (2) all technical lemmas in Appendix E hold when the Gaussian sketching $\mathbf{S}$ is replaced by a row-wise sub-Gaussian matrix by concentration bounds on quadratic forms of sub-Gaussian vectors (e.g., Theorem 1 in Hsu et al. (2012)), and on sub-Gaussian covariance matrices (e.g., Example 1.5 in Zhivotovskiy (2024)); (3) the norm of sketched samples satisfy the same concentration bounds by e.g., Theorem 6.5 in Wainwright (2019).

On the other hand, for the lower bounds, the Gaussian assumption is still required in order to establish the conditional independence of $\tilde{\epsilon}_i = y_i - \mathbf{x}_i^\top \mathbf{S}^\top \mathbf{v}^*$ and $\mathbf{S}\mathbf{x}_i$ given $(\mathbf{S}, \mathbf{w}^*)$ in the proof of Theorem 3.1 and Lemma D.4.

In addition, we also conduct experiments to check our justification of (b). We generate data $\mathbf{x} = (x_1, \ldots, x_d)^\top$ from the distribution where x_i are independent and

$$\mathbb{P}(x_i = 1) = \mathbb{P}(x_i = -1) = i^{-a}/c_0, \quad \mathbb{P}(x_i = 0) = 1 - 2 \cdot \mathbb{P}(x_i = 1),$$

with $a = 2, b = 1.5$ and $c_0 = 2\sum_{i=1}^d i^{-a}$. Note that $\mathbf{x}$ satisfies Assumption 1A' but not Assumption 1A when $d = \infty$. We run the experiment under the same setting and choice of hyperparameters as in Figure 1(a). Similar to the Gaussian case, in Figure 2, we observe that the excess test error of one-pass SGD and multi-pass SGD both exhibit power-law scaling in the number of effective samples $N_{\texttt{eff}}$. Moreover, the fitted slopes are both close to the theoretical prediction in Corollary 3.3 ($0.34 \approx 0.33 = (1-b)/b$ for multi-pass SGD and $0.26 \approx 0.25 = (1-b)/a$ for one-pass SGD).

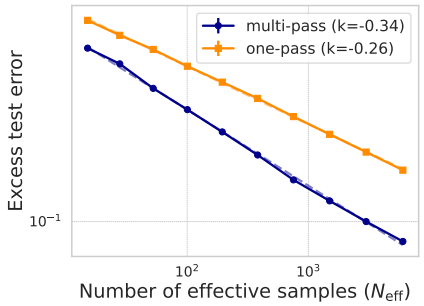


Figure 2: Multi-pass SGD versus one-pass SGD for non-Gaussian design. Multi-pass SGD is run for $L \eqsim N^{a/b}$ steps. We use linear functions to fit the excess test error in log-log scale. The fitted exponents (k) are close to the theoretical predictions in Corollary 3.3. The errorbars denote the ± 1 standard deviation of the expected excess test error over 100 i.i.d. samples of $(\mathbf{S}, \mathbf{w}^*)$. Parameters: $\sigma^2 = 1$, $d = 10000$, $M = 1000$, $\gamma = 0.1$.

B Bias error

B.1 An upper bound

Lemma B.1 (An upper bound on the GD bias term)**.** *Suppose $L_{\mathtt{eff}} \lesssim N^a/\gamma$ and Assumption 1A and 3 hold. Under the notation in Theorem 3.1 and its proof, for any $\mathbf{w}^* \in \mathbb{H}$ and $k \leqslant M/3$ such that $r(\mathbf{H}) \geqslant k + M$, the bias term*

$$\begin{aligned}\mathsf{Bias}(\mathbf{w}^*) &= \mathbb{E}_{\mathbf{X}}\Big\| \prod_{t=1}^{L} \Big(\mathbf{I} - \gamma_t \widehat{\mathbf{\Sigma}}\Big)\mathbf{v}^* \Big\|_{\mathbf{\Sigma}}^2 \\ &\leqslant \frac{c\|\mathbf{w}^*_{0:k}\|_2^2}{L_{\mathtt{eff}}\gamma} \cdot \left[\frac{\mu_{M/2}(\mathbf{A}_k)}{\mu_M(\mathbf{A}_k)}\right]^2 + B_{\mathbf{B}} \cdot \|\mathbf{w}^*_{k:\infty}\|^2_{\mathbf{H}_{k:\infty}}\end{aligned}$$

with probability at least $1 - e^{-\Omega(M)}$ over the randomness of $\mathbf{S}$, where $\mathbf{A}_k = \mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}$ and

$$B_{\mathbf{B}} := c\left(1 + [(L_{\mathtt{eff}}\gamma)^2 + 1]\Big(\frac{\mathrm{tr}^2(\mathbf{\Sigma}_{\tilde{k}:\infty})}{N^2} + \|\mathbf{\Sigma}_{\tilde{k}:\infty}\|_2^2 + \frac{\mathrm{tr}(\mathbf{\Sigma}^2_{\tilde{k}:\infty})}{N} + \sqrt{\frac{\mathrm{tr}(\mathbf{\Sigma}^4_{\tilde{k}:\infty})}{N}}\Big)\right)$$

for some constant $c > 0$ and $\tilde{k} = \lfloor N/2 \rfloor$.[3]

Proof of Lemma B.1. Without loss of generality, we assume the covariance matrix $\mathbf{H} = \mathrm{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_d\}$ where $\lambda_i \geqslant \lambda_j$ for any $i \geqslant j$. Let $(\tilde{\lambda}_1, \tilde{\lambda}_2, \ldots, \tilde{\lambda}_M)$ denote the eigenvalues of $\mathbf{\Sigma}$ in non-increasing order. Moreover, we introduce $\mathbf{z}_1, \ldots \mathbf{z}_N \overset{iid}{\sim} \mathcal{N}(0, \mathbf{I}_M/N)$ and write $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_N)^\top$. It can be verified that $\mathbf{X}\mathbf{S}^\top/\sqrt{N} \overset{d}{=} \mathbf{Z}\mathbf{\Sigma}^{1/2}$ conditioned on $\mathbf{S}$. Throughout the proof, by a union bound argument, we w.l.o.g. assume the conditions (1), (2), (3) and (4) in Assumption 3 always hold.

Define $\mathsf{M} := \prod_{t=1}^{L}(\mathbf{I} - \gamma_t\widehat{\mathbf{\Sigma}})\mathbf{\Sigma}\prod_{t=1}^{L}(\mathbf{I} - \gamma_t\widehat{\mathbf{\Sigma}})$ and recall that $\mathbf{v}^* = \big(\mathbf{S}\mathbf{H}\mathbf{S}^\top\big)^{-1}\mathbf{S}\mathbf{H}\mathbf{w}^*$. Substituting

$$\mathbf{S}\mathbf{H} = (\mathbf{S}_{0:k}\mathbf{H}_{0:k} \quad \mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty})$$

into $\mathbf{v}^*$, we have

$$\begin{aligned}\mathsf{Bias}(\mathbf{w}^*) &= \mathbb{E}_{\mathbf{X}}[\mathbf{v}^{*\top}\mathsf{M}\mathbf{v}^*] \\ &= \mathbb{E}_{\mathbf{X}}[\mathbf{w}^{*\top}\mathbf{H}\mathbf{S}^\top(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathsf{M}(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathbf{S}\mathbf{H}\mathbf{w}^*] \\ &\leqslant 2T_1 + 2T_2,\end{aligned}$$

[3] If $\tilde{k} > M$ then $\mathbf{\Sigma}_{\tilde{k}:\infty} := \mathbf{0}$.

where

$$T_1 := \mathbb{E}_{\mathbf{X}}[(\mathbf{w}^*_{0:k})^\top \mathbf{H}_{0:k}\mathbf{S}_{0:k}^\top(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathsf{M}(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathbf{S}_{0:k}\mathbf{H}_{0:k}\mathbf{w}^*_{0:k}], \tag{5}$$

$$T_2 := \mathbb{E}_{\mathbf{X}}[(\mathbf{w}^*_{k:\infty})^\top \mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathsf{M}(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{w}^*_{k:\infty}]. \tag{6}$$

We will show the following results at the end of the proof. First, with probability $1 - e^{-\Omega(M)}$

$$T_1 \leqslant \frac{c\|\mathbf{w}^*_{0:k}\|_2^2}{L_{\mathtt{eff}}\gamma} \cdot \left[\frac{\mu_{M/2}(\mathbf{A}_k)}{\mu_M(\mathbf{A}_k)}\right]^2 \tag{7a}$$

for some constant $c > 0$. Moreover,

$$T_2 \leqslant B_{\mathbf{B}} \cdot \|\mathbf{w}^*_{k:\infty}\|^2_{\mathbf{H}_{k:\infty}}. \tag{7b}$$

Combining Eq. (7a) and (7b) gives Lemma B.1.

Proof of claim (7a)**.** By definition of T_1, we have

$$\begin{aligned} T_1 &\leqslant \|\mathbf{H}_{0:k}\mathbf{S}_{0:k}^\top(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathsf{M}(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathbf{S}_{0:k}\mathbf{H}_{0:k}\|_2 \cdot \|\mathbf{w}^*_{0:k}\|_2^2 \\ &\leqslant \|\mathsf{M}\|_2 \cdot \|(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathbf{S}_{0:k}\mathbf{H}_{0:k}\|_2^2 \cdot \|\mathbf{w}^*_{0:k}\|_2^2. \end{aligned}$$

for some constant $c > 0$. By Eq. (23) in the proof of Lemma D.1 in Lin et al. (2024), we have

$$\|(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathbf{S}_{0:k}\mathbf{H}_{0:k}\|_2 \leqslant c \cdot \frac{\mu_{M/2}(\mathbf{A}_k)}{\mu_M(\mathbf{A}_k)} \tag{8}$$

for some constant $c > 0$ with probability at least $1 - e^{-\Omega(M)}$. Thus, it remains to show

$$\mathbb{E}_{\mathbf{X}}[\|\mathsf{M}\|_2] \leqslant \frac{c}{L_{\mathtt{eff}}\gamma} \tag{9}$$

for some constant $c > 0$.

Let $\lambda > 0$ be a fixed value to be specified later. Note that

$$\begin{aligned} \|\mathsf{M}\|_2 = \Big\|\prod_{t=1}^{L}(\mathbf{I} - \gamma_t\widehat{\mathbf{\Sigma}})\mathbf{\Sigma}^{1/2}\Big\|_2^2 &= \Big\|\prod_{t=1}^{L}(\mathbf{I} - \gamma_t\widehat{\mathbf{\Sigma}})(\widehat{\mathbf{\Sigma}} + \lambda\mathbf{I}_M)^{1/2}(\widehat{\mathbf{\Sigma}} + \lambda\mathbf{I}_M)^{-1/2}\mathbf{\Sigma}^{1/2}\Big\|_2^2 \\ &\leqslant \Big\|\prod_{t=1}^{L}(\mathbf{I} - \gamma_t\widehat{\mathbf{\Sigma}})(\widehat{\mathbf{\Sigma}} + \lambda\mathbf{I}_M)^{1/2}\Big\|_2^2 \cdot \|(\widehat{\mathbf{\Sigma}} + \lambda\mathbf{I}_M)^{-1/2}\mathbf{\Sigma}^{1/2}\|_2^2 \\ &\overset{(i)}{\leqslant} \Big(\Big\|\prod_{t=1}^{L}(\mathbf{I} - \gamma_t\widehat{\mathbf{\Sigma}})^2\widehat{\mathbf{\Sigma}}\Big\|_2 + \lambda\Big) \cdot \|(\widehat{\mathbf{\Sigma}} + \lambda\mathbf{I}_M)^{-1/2}\mathbf{\Sigma}^{1/2}\|_2^2 \\ &\overset{(ii)}{\leqslant} \Big(\frac{c}{L_{\mathtt{eff}}\gamma_0} + \lambda\Big) \cdot \|(\widehat{\mathbf{\Sigma}} + \lambda\mathbf{I}_M)^{-1/2}\mathbf{\Sigma}^{1/2}\|_2^2, \end{aligned}$$

where step (i) uses the fact that $\|\mathbf{I} - \gamma_t\widehat{\mathbf{\Sigma}}\|_2 \leqslant 1$ by the stepsize assumption and step (ii) follows from the stepsize assumption (2) that $\gamma_t = \gamma_0$ for $t \in [L_{\mathtt{eff}}]$, combined with the fact that $\sup_{x\in[0,1/\gamma_0]} x(1-\gamma_0 x)^{2L_{\mathtt{eff}}} \leqslant c/(\gamma_0 L_{\mathtt{eff}})$ for some constant $c > 0$.

Recall that we assume $\mathbb{P}(\gamma_0 < \gamma/t) \leqslant N^{-ct}$ for some constant $c > 0$ and all $t \geqslant 1$. Thus, Eq. (9) follows immediately from choosing $\lambda = 1/(L_{\mathtt{eff}}\gamma)$ in the last display, applying Cauchy-Schwartz inequality and Lemma E.1, and noting that $(1 + {L_{\mathtt{eff}}}^2\gamma^2\exp(-cN)) \lesssim 1$ for $L_{\mathtt{eff}} \lesssim N^a/\gamma$.

Proof of claim (7b)**.** Let $\mathbf{B} = \mathbb{E}_{\mathbf{X}}\Big[\mathbf{\Sigma}^{-1/2}\prod_{t=1}^{L}(\mathbf{I}-\gamma_t\widehat{\mathbf{\Sigma}})\mathbf{\Sigma}\prod_{t=1}^{L}(\mathbf{I}-\gamma_t\widehat{\mathbf{\Sigma}})\mathbf{\Sigma}^{-1/2}\Big]$. By definition of T_2 in Eq. (6), we have

$$\begin{aligned} T_2 &= {\mathbf{w}^*_{k:\infty}}^\top \mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top\mathbf{\Sigma}^{-1/2}\mathbf{B}\mathbf{\Sigma}^{-1/2}\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{w}^*_{k:\infty} \\ &\leqslant \|\mathbf{B}\|_2 \cdot \|\mathbf{H}^{1/2}_{k:\infty}\mathbf{S}_{k:\infty}^\top\mathbf{\Sigma}^{-1}\mathbf{S}_{k:\infty}\mathbf{H}^{1/2}_{k:\infty}\| \cdot \|\mathbf{w}^*_{k:\infty}\|^2_{\mathbf{H}_{k:\infty}} \end{aligned}$$

$$\leqslant \|\mathbf{B}\|_2 \cdot \|\mathbf{w}^*_{k:\infty}\|^2_{\mathbf{H}_{k:\infty}},$$

where the last line follows since

$$\begin{aligned}\|\mathbf{H}^{1/2}_{k:\infty}\mathbf{S}^\top_{k:\infty}\boldsymbol{\Sigma}^{-1}\mathbf{S}_{k:\infty}\mathbf{H}^{1/2}_{k:\infty}\|_2 &= \|\mathbf{H}^{1/2}_{k:\infty}\mathbf{S}^\top_{k:\infty}(\mathbf{S}_{0:k}\mathbf{H}_{0:k}\mathbf{S}^\top_{0:k} + \mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}^\top_{k:\infty})^{-1}\mathbf{S}_{k:\infty}\mathbf{H}^{1/2}_{k:\infty}\|_2 \\ &\leqslant \|\mathbf{H}^{1/2}_{k:\infty}\mathbf{S}^\top_{k:\infty}\mathbf{A}_k^{-1}\mathbf{S}_{k:\infty}\mathbf{H}^{1/2}_{k:\infty}\|_2 \leqslant 1.\end{aligned}$$

In the following, we will show that $\|\mathbf{B}\|_2 \leqslant B_{\mathbf{B}}$, which immediately yields claim (7b).

Let $\overline{\boldsymbol{\Sigma}} = \mathbf{X}\mathbf{S}^\top\mathbf{S}\mathbf{X}^\top/N$. To compute $\|\mathbf{B}\|_2$, note that

$$\begin{aligned}\prod_{t=1}^{L}(\mathbf{I}-\gamma_t\widehat{\boldsymbol{\Sigma}}) &= \prod_{t=1}^{L-1}(\mathbf{I}-\gamma_t\widehat{\boldsymbol{\Sigma}}) - \gamma_L\widehat{\boldsymbol{\Sigma}}\prod_{t=1}^{L-1}(\mathbf{I}-\gamma_t\widehat{\boldsymbol{\Sigma}}) \\ &\overset{(iii)}{=} \prod_{t=1}^{L-1}(\mathbf{I}-\gamma_t\widehat{\boldsymbol{\Sigma}}) - \frac{1}{N}\mathbf{S}\mathbf{X}^\top\Big[\gamma_L\prod_{t=1}^{L-1}(\mathbf{I}-\gamma_t\overline{\boldsymbol{\Sigma}})\Big]\mathbf{X}\mathbf{S}^\top \\ &= \mathbf{I} - \frac{1}{N}\mathbf{S}\mathbf{X}^\top\Big[\sum_{i=0}^{L-1}\gamma_{i+1}\prod_{t=1}^{i}(\mathbf{I}-\gamma_t\overline{\boldsymbol{\Sigma}})\Big]\mathbf{X}\mathbf{S}^\top =: \mathbf{I}-\mathbf{C},\end{aligned}$$

where step (iii) uses $\mathbf{X}\mathbf{S}^\top(\mathbf{I}-\gamma_L\widehat{\boldsymbol{\Sigma}}) = (\mathbf{I}-\gamma_L\overline{\boldsymbol{\Sigma}})\mathbf{X}\mathbf{S}^\top$. Recall that $\mathbf{X}\mathbf{S}^\top/\sqrt{N} \overset{d}{=} \mathbf{Z}\boldsymbol{\Sigma}^{1/2}$ conditioned on $\mathbf{S}$, we can thus rewrite

$$\begin{aligned}\mathbf{B} &= \mathbb{E}_{\mathbf{Z}}[\boldsymbol{\Sigma}^{-1/2}(\mathbf{I}-\mathbf{C})\boldsymbol{\Sigma}(\mathbf{I}-\mathbf{C})\boldsymbol{\Sigma}^{-1/2}] \preceq 2\mathbf{I} + 2\mathbb{E}[\boldsymbol{\Sigma}^{-1/2}\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}\boldsymbol{\Sigma}^{-1/2}] \\ &= 2\mathbf{I} + 2\mathbb{E}_{\mathbf{Z}}\Big[\mathbf{Z}^\top\Big[\sum_{i=0}^{L-1}\gamma_{i+1}\prod_{t=1}^{i}(\mathbf{I}-\gamma_t\overline{\boldsymbol{\Sigma}})\Big]\mathbf{Z}\boldsymbol{\Sigma}^2\mathbf{Z}^\top\Big[\sum_{i=0}^{L-1}\gamma_{i+1}\prod_{t=1}^{i}(\mathbf{I}-\gamma_t\overline{\boldsymbol{\Sigma}})\Big]\mathbf{Z}\Big], \text{ where } \overline{\boldsymbol{\Sigma}} = \mathbf{Z}\boldsymbol{\Sigma}\mathbf{Z}^\top.\end{aligned}$$

Introduce the shorthand $\mathbf{R}_1 = \sum_{i=0}^{L-1}\gamma_{i+1}\prod_{t=1}^{i}(\mathbf{I}-\gamma_t\overline{\boldsymbol{\Sigma}})$ and $\mathbf{R}_1(k) := (\mathbf{I} - (\mathbf{I} - \gamma_{kL_{\mathtt{eff}}+1}\overline{\boldsymbol{\Sigma}})^{L_{\mathtt{eff}}})/\overline{\boldsymbol{\Sigma}}$ for $k \in [0, \lfloor \log L_{\mathtt{eff}} \rfloor - 1]$. Note that $\|\mathbf{R}_1(k)\|_2 \leqslant L_{\mathtt{eff}}\cdot\gamma_{kL_{\mathtt{eff}}+1}$ since $\sup_{x\in[0,1/\gamma_{kL_{\mathtt{eff}}+1}]}[(1-(1-\gamma_{kL_{\mathtt{eff}}+1}x)^{L_{\mathtt{eff}}})/x] = L_{\mathtt{eff}}\cdot\gamma_{kL_{\mathtt{eff}}+1}$. Therefore

$$\|\mathbf{R}_1\|_2 = \Big\|\sum_{i=0}^{L-1}\gamma_{i+1}\prod_{t=1}^{i}(\mathbf{I}-\gamma_t\overline{\boldsymbol{\Sigma}})\Big\|_2 \leqslant \Big\|\sum_{k=0}^{\lfloor\log L_{\mathtt{eff}}\rfloor-1}\mathbf{R}_1(k)\Big\|_2 \leqslant \sum_{k=0}^{\lfloor\log L_{\mathtt{eff}}\rfloor-1} L_{\mathtt{eff}}\cdot\gamma_{kL_{\mathtt{eff}}+1} \leqslant 2L_{\mathtt{eff}}\gamma,$$

where the last inequality follows from (2) and the definition of γ_0.

We consider two cases.

Case 1: $M \leqslant N/2$. In this case, we have

$$\mathbf{Z}\boldsymbol{\Sigma}^2\mathbf{Z}^\top \preceq 5\cdot(\mathbf{Z}\boldsymbol{\Sigma}\mathbf{Z}^\top)^2 \tag{10}$$

with probability at least $1-e^{-\Omega(N)}$ since $\mathbb{P}(\mathbf{Z}^\top\mathbf{Z} \succeq \mathbf{I}_M/5) \geqslant 1-e^{-\Omega(N)}$ by concentration of Guassian covariance matrix (see e.g. Theorem 6.1 in Wainwright (2019)). Moreover, since $\mathrm{tr}(\boldsymbol{\Sigma}^2) \lesssim 1$,

$$\begin{aligned}\mathbf{Z}^\top\mathbf{R}_1\mathbf{Z}\boldsymbol{\Sigma}^2\mathbf{Z}^\top\mathbf{R}_1\mathbf{Z} &\preceq c\cdot\mathbf{Z}^\top\mathbf{R}_1\mathbf{Z}\mathbf{Z}^\top\mathbf{R}_1\mathbf{Z} \preceq c\|\mathbf{Z}^\top\mathbf{R}_1\mathbf{Z}\|_2^2\cdot\mathbf{I}_M \\ &\preceq c(L_{\mathtt{eff}}\gamma)^2\|\mathbf{Z}^\top\mathbf{Z}\|_2^2\cdot\mathbf{I}_M.\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{E}_{\mathbf{Z}}[\mathbf{Z}^\top\mathbf{R}_1\mathbf{Z}\boldsymbol{\Sigma}^2\mathbf{Z}^\top\mathbf{R}_1\mathbf{Z}] &\preceq c\mathbb{E}_{\mathbf{Z}}[\mathbf{Z}^\top\mathbf{R}_1\overline{\boldsymbol{\Sigma}}^2\mathbf{R}_1\mathbf{Z}] + c\mathbb{E}[(L_{\mathtt{eff}}\gamma)^2\|\mathbf{Z}^\top\mathbf{Z}\|_2^2\mathbf{1}_{\{\mathbf{Z}^\top\mathbf{Z}\not\succeq\mathbf{I}_M/5\}}]\cdot\mathbf{I}_M \\ &\preceq c\mathbb{E}_{\mathbf{Z}}[\mathbf{Z}^\top\mathbf{Z}] + c(L_{\mathtt{eff}}\gamma)^2\exp(-c'N)\cdot\mathbf{I}_M \preceq c\mathbf{I}_M\end{aligned} \tag{11}$$

for some constant $c, c' > 0$, where the second line uses the fact that $\|\mathbf{R}_1\overline{\boldsymbol{\Sigma}}\|_2 = \|\mathbf{I} - \prod_{t=1}^{L}(\mathbf{I}-\gamma_t\overline{\boldsymbol{\Sigma}})\|_2 \leqslant 1$, concentration properties of the empirical covariance matrix $\mathbf{Z}^\top\mathbf{Z}$, and $\mathbb{E}_{\mathbf{Z}}[\mathbf{Z}^\top\mathbf{Z}] = \mathbf{I}_M$. As a result, $\|\mathbf{B}\|_2 \leqslant 2 + 2\|\mathbb{E}_{\mathbf{Z}}[\mathbf{Z}^\top\mathbf{R}_1\mathbf{Z}\boldsymbol{\Sigma}^2\mathbf{Z}^\top\mathbf{R}_1\mathbf{Z}]\|_2 \lesssim 1$.

Case2: $M > N/2$. Let $\tilde{k} = N/2$. W.l.o.g., we assume $\mathbf{\Sigma}$ is a diagonal matrix with eigenvalues $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_M$ in non-increasing order. With probability at least $1 - e^{-\Omega(N)}$, we have the decomposition

$$\begin{aligned}
\mathbf{Z}^\top \mathbf{R}_1 \mathbf{Z} \mathbf{\Sigma}^2 \mathbf{Z}^\top \mathbf{R}_1 \mathbf{Z} &\preceq 2\mathbf{Z}^\top \mathbf{R}_1 (\mathbf{Z}_{0:\tilde{k}} \mathbf{\Sigma}_{0:\tilde{k}}^2 \mathbf{Z}_{0:\tilde{k}}^\top) \mathbf{R}_1 \mathbf{Z} + 2\mathbf{Z}^\top \mathbf{R}_1 (\mathbf{Z}_{\tilde{k}:\infty} \mathbf{\Sigma}_{\tilde{k}:\infty}^2 \mathbf{Z}_{\tilde{k}:\infty}^\top) \mathbf{R}_1 \mathbf{Z} \\
&\preceq c\mathbf{Z}^\top \mathbf{R}_1 (\mathbf{Z}_{0:\tilde{k}} \mathbf{\Sigma}_{0:\tilde{k}} \mathbf{Z}_{0:\tilde{k}}^\top)^2 \mathbf{R}_1 \mathbf{Z} + 2\mathbf{Z}^\top \mathbf{R}_1 (\mathbf{Z}_{\tilde{k}:\infty} \mathbf{\Sigma}_{\tilde{k}:\infty}^2 \mathbf{Z}_{\tilde{k}:\infty}^\top) \mathbf{R}_1 \mathbf{Z} \\
&\preceq c(\|\mathbf{R}_1 (\mathbf{Z}_{0:\tilde{k}} \mathbf{\Sigma}_{0:\tilde{k}} \mathbf{Z}_{0:\tilde{k}}^\top)\|_2^2 + \|\mathbf{Z}_{\tilde{k}:\infty} \mathbf{\Sigma}_{\tilde{k}:\infty}^2 \mathbf{Z}_{\tilde{k}:\infty}^\top\|_2) \mathbf{Z}^\top \mathbf{Z} \\
&\preceq c(\|\mathbf{R}_1 (\mathbf{Z}_{\tilde{k}:\infty} \mathbf{\Sigma}_{\tilde{k}:\infty} \mathbf{Z}_{\tilde{k}:\infty}^\top)\|_2^2 + \|\mathbf{R}_1 \overline{\mathbf{\Sigma}}\|_2^2 + \|\mathbf{Z}_{\tilde{k}:\infty} \mathbf{\Sigma}_{\tilde{k}:\infty}^2 \mathbf{Z}_{\tilde{k}:\infty}^\top\|_2) \mathbf{Z}^\top \mathbf{Z},
\end{aligned}$$

where the second line use $\mathbf{Z}_{0:\tilde{k}}^\top \mathbf{Z}_{0:\tilde{k}} \succeq \mathbf{I}_k/5$ with probability at least $1 - e^{-\Omega(N)}$, the last line follows from a triangle inequality. Since $\|\mathbf{R}_1\|_2 \leqslant L_{\mathtt{eff}}\gamma$ and $\|\mathbf{R}_1 \overline{\mathbf{\Sigma}}\|_2 = \|\mathbf{I} - \prod_{t=1}^L (\mathbf{I} - \gamma_t \overline{\mathbf{\Sigma}})\|_2 \leqslant 1$, continuing the calculation, we obtain

$$\mathbf{Z}^\top \mathbf{R}_1 \mathbf{Z} \mathbf{\Sigma}^2 \mathbf{Z}^\top \mathbf{R}_1 \mathbf{Z} \preceq c\big((L_{\mathtt{eff}}\gamma)^2 \|\mathbf{Z}_{\tilde{k}:\infty} \mathbf{\Sigma}_{\tilde{k}:\infty} \mathbf{Z}_{\tilde{k}:\infty}^\top\|_2^2 + 1 + \|\mathbf{Z}_{\tilde{k}:\infty} \mathbf{\Sigma}_{\tilde{k}:\infty}^2 \mathbf{Z}_{\tilde{k}:\infty}^\top\|_2\big) \mathbf{Z}^\top \mathbf{Z}. \tag{12}$$

Since we have by Lemma E.3 that

$$\begin{aligned}
\|\mathbf{Z}_{\tilde{k}:\infty} \mathbf{\Sigma}_{\tilde{k}:\infty} \mathbf{Z}_{\tilde{k}:\infty}^\top\|_2 \leqslant c\left(\frac{\mathrm{tr}(\mathbf{\Sigma}_{\tilde{k}:\infty})}{N} + \|\mathbf{\Sigma}_{\tilde{k}:\infty}\|_2 + \sqrt{\frac{\mathrm{tr}(\mathbf{\Sigma}_{\tilde{k}:\infty}^2)}{N}}\right) \\
+ c\left(\frac{\|\mathbf{\Sigma}_{\tilde{k}:\infty}\|_2}{N} \log\frac{1}{\delta} + \frac{\sqrt{\mathrm{tr}(\mathbf{\Sigma}_{\tilde{k}:\infty}^2)\log(1/\delta)}}{N}\right)
\end{aligned} \tag{13}$$

with probability at least $1-\delta$, and $\mathrm{tr}(\mathbf{\Sigma}_{\tilde{k}:\infty}^2) \leqslant \mathrm{tr}(\mathbf{\Sigma}^2) \lesssim 1$, it can be verified by a standard truncation argument that

$$\mathbb{E}[\|\mathbf{Z}_{\tilde{k}:\infty} \mathbf{\Sigma}_{\tilde{k}:\infty} \mathbf{Z}_{\tilde{k}:\infty}^\top\|_2^2 \cdot \mathbf{Z}^\top \mathbf{Z}] \preceq c\left(\frac{\mathrm{tr}^2(\mathbf{\Sigma}_{\tilde{k}:\infty})}{N^2} + \|\mathbf{\Sigma}_{\tilde{k}:\infty}\|_2^2 + \frac{\mathrm{tr}(\mathbf{\Sigma}_{\tilde{k}:\infty}^2)}{N}\right) \cdot \mathbf{I}_M.$$

A similar bound can be established for $\mathbb{E}[\|\mathbf{Z}_{\tilde{k}:\infty} \mathbf{\Sigma}_{\tilde{k}:\infty}^2 \mathbf{Z}_{\tilde{k}:\infty}^\top\|_2]$. Finally, substituting the bounds on the expectations into Eq. (12), we obtain

$$\begin{aligned}
\|\mathbf{B}\|_2 &\leqslant 2 + 2\|\mathbb{E}[\mathbf{Z}^\top \mathbf{R}_1 \mathbf{Z} \mathbf{\Sigma}^2 \mathbf{Z}^\top \mathbf{R}_1 \mathbf{Z}]\|_2 \\
&\leqslant c\left(1 + [(L_{\mathtt{eff}}\gamma)^2 + 1]\Big(\frac{\mathrm{tr}^2(\mathbf{\Sigma}_{\tilde{k}:\infty})}{N^2} + \|\mathbf{\Sigma}_{\tilde{k}:\infty}\|_2^2 + \frac{\mathrm{tr}(\mathbf{\Sigma}_{\tilde{k}:\infty}^2)}{N} + \sqrt{\frac{\mathrm{tr}(\mathbf{\Sigma}_{\tilde{k}:\infty}^4)}{N}}\Big)\right).
\end{aligned}$$

□

B.2 A lower bound

Lemma B.2 (A lower bound on the GD bias term)**.** *Let Assumption 1A and 3 hold. Define* $\mathbf{H}^{\mathbf{w}} := \mathbb{E}[\mathbf{w}^* \mathbf{w}^{*\top}]$ *and* $\mathbf{\Sigma}_{\mathbf{w}} := \mathbf{S}\mathbf{H}\mathbf{H}^{\mathbf{w}}\mathbf{H}\mathbf{S}^\top$*. Under the notation in Theorem 3.1 and its proof, the bias term satisfies*

$$\begin{aligned}
\mathbb{E}_{\mathbf{w}^*}[\mathsf{Bias}(\mathbf{w}^*)] &= \mathbb{E}_{\mathbf{w}^*} \mathbb{E}\Big\| \prod_{t=1}^L \Big(\mathbf{I} - \gamma_t \widehat{\mathbf{\Sigma}}\Big) \mathbf{v}^* \Big\|_{\mathbf{\Sigma}}^2 \\
&\gtrsim \sum_{i=2\bar{t}+1}^M \frac{\mu_{3i}(\mathbf{\Sigma}_{\mathbf{w}})}{\mu_i(\mathbf{\Sigma})}
\end{aligned}$$

with probability at least $1 - e^{-\Omega(M)}$*, where* $\bar{t} := \mathbb{E}_{\mathbf{X}}[\#\{i \in [M] : \widehat{\lambda}_i L_{\mathtt{eff}}\gamma_0 > 1/4\}]$ *and* $(\widehat{\lambda}_i)_{i=1}^M$ *are the eigenvalues of* $\widehat{\mathbf{\Sigma}}$*.*

Proof of Lemma B.2. Similar to the proof of Lemma B.1, w.l.o.g., we assume the covariance matrix $\mathbf{H} = \mathrm{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_d\}$ where $\lambda_i \geqslant \lambda_j$ for any $i \geqslant j$. Let $(\tilde{\lambda}_1, \tilde{\lambda}_2, \ldots, \tilde{\lambda}_M)$ denote the eigenvalues of $\mathbf{\Sigma}$ in non-increasing order. Moreover, we introduce $\mathbf{z}_1, \ldots \mathbf{z}_N \overset{iid}{\sim} \mathcal{N}(0, \mathbf{I}_M/N)$ and write $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_N)^\top$. It can be shown that $\mathbf{X}\mathbf{S}^\top/\sqrt{N} \overset{d}{=} \mathbf{Z}\mathbf{\Sigma}^{1/2}$ conditioned on $\mathbf{S}$.

Let $\mathsf{C}_L := \prod_{t=1}^{L}(\mathbf{I} - \gamma_t \widehat{\mathbf{\Sigma}})$. By definition,

$$\mathsf{Bias}(\mathbf{w}^*) = \mathbb{E}_{\mathbf{X}}\Big\| \prod_{t=1}^{L}\big(\mathbf{I} - \gamma_t \widehat{\mathbf{\Sigma}}\big)\mathbf{v}^* \Big\|_{\mathbf{\Sigma}}^2 = \mathbf{v}^{*\top}\mathbb{E}_{\mathbf{X}}[\mathsf{C}_L \mathbf{\Sigma} \mathsf{C}_L]\mathbf{v}^*. \tag{14}$$

Adopt the shorthand $\mathbf{\Sigma}_{\mathbf{w}}$ for $\mathbf{S}\mathbf{H}\mathbf{H}^{\mathbf{w}}\mathbf{H}\mathbf{S}^\top$. Substituting the definition of $\mathbf{v}^*$ into the expression and noting that $\mathbb{E}[\mathbf{w}^*\mathbf{w}^{*\top}] = \mathbf{H}^{\mathbf{w}}$, we have

$$\begin{aligned}
\mathbb{E}_{\mathbf{w}^*}[\mathsf{Bias}(\mathbf{w}^*)] &= \mathbb{E}_{\mathbf{w}^*}[\mathbf{v}^{*\top}\mathbb{E}_{\mathbf{X}}[\mathsf{C}_L \mathbf{\Sigma} \mathsf{C}_L]\mathbf{v}^*] \\
&= \mathrm{tr}(\mathbf{H}\mathbf{S}^\top(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathbb{E}_{\mathbf{X}}[\mathsf{C}_L \mathbf{\Sigma} \mathsf{C}_L](\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathbf{S}\mathbf{H}\mathbf{H}^{\mathbf{w}}) \\
&\overset{(i)}{\geqslant} \mathrm{tr}(\mathbf{H}\mathbf{S}^\top(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathbb{E}_{\mathbf{X}}[\mathsf{C}_L]\mathbf{\Sigma}\mathbb{E}_{\mathbf{X}}[\mathsf{C}_L^\top](\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathbf{S}\mathbf{H}\mathbf{H}^{\mathbf{w}}) \\
&= \mathbb{E}_{\mathbf{X}}[\mathrm{tr}(\mathbf{\Sigma}^{-1/2}\mathbb{E}_{\mathbf{X}}[\mathsf{C}_L]\mathbf{\Sigma}\mathbb{E}_{\mathbf{X}}[\mathsf{C}_L^\top]\mathbf{\Sigma}^{-1/2}\mathbf{\Sigma}^{-1/2}\mathbf{\Sigma}_{\mathbf{w}}\mathbf{\Sigma}^{-1/2})],
\end{aligned}$$

where step (i) uses the fact that $\mathbb{E}[\mathbf{Y}]\mathbb{E}[\mathbf{Y}^\top] \preceq \mathbb{E}[\mathbf{Y}\mathbf{Y}^\top]$ for any random matrix $\mathbf{Y}$. We claim that

$$\mathbf{\Sigma}^{-1/2}\mathbb{E}_{\mathbf{X}}[\mathsf{C}_L]\mathbf{\Sigma}^{1/2} = \mathbb{E}_{\mathbf{X}}[\mathsf{C}_L], \quad \text{and} \tag{15a}$$

$$\mu_{M-i+1}(\mathbb{E}_{\mathbf{X}}[\mathsf{C}_L]) \geqslant \frac{1}{2e} \tag{15b}$$

for all $i \in [2\bar{t}+1, M]$, where $\bar{t} := \mathbb{E}_{\mathbf{X}}[\#\{i \in [M] : \widehat{\lambda}_i L \gamma_0 > 1/4\}]$.

The proof of these two claims will be given momentarily. Continuing the calculation using the claims and Von Neumann's trace inequality, we obtain

$$\begin{aligned}
\mathbb{E}_{\mathbf{w}^*}\mathsf{Bias}(\mathbf{w}^*) &\geqslant \mathbb{E}_{\mathbf{X}}[\mathrm{tr}(\mathbb{E}_{\mathbf{X}}[\mathsf{C}_L]^2\mathbf{\Sigma}^{-1/2}\mathbf{\Sigma}_{\mathbf{w}}\mathbf{\Sigma}^{-1/2})] \\
&\geqslant \sum_{i=1}^{M} \mu_{M-i+1}^2(\mathbb{E}_{\mathbf{X}}[\mathsf{C}_L]) \cdot \mu_i(\mathbf{\Sigma}^{-1/2}\mathbf{\Sigma}_{\mathbf{w}}\mathbf{\Sigma}^{-1/2}) \\
&\geqslant \sum_{i=2\bar{t}+1}^{M} \mu_{M-i+1}^2(\mathbb{E}_{\mathbf{X}}[\mathsf{C}_L]) \cdot \mu_i(\mathbf{\Sigma}^{-1/2}\mathbf{\Sigma}_{\mathbf{w}}\mathbf{\Sigma}^{-1/2}) \gtrsim \sum_{i=2\bar{t}+1}^{M} \mu_i(\mathbf{\Sigma}^{-1/2}\mathbf{\Sigma}_{\mathbf{w}}\mathbf{\Sigma}^{-1/2}).
\end{aligned}$$

Since $\mu_{i+j+1}(XY) \leqslant \mu_{i+1}(X)\mu_{j+1}(Y)$ for all i, j and any matrices X, Y of matching dimensions, we have

$$\mu_i(\mathbf{\Sigma}^{-1/2}\mathbf{\Sigma}_{\mathbf{w}}\mathbf{\Sigma}^{-1/2}) \geqslant \frac{\mu_{2i-1}(\mathbf{\Sigma}_{\mathbf{w}}\mathbf{\Sigma}^{-1/2})}{\mu_i(\mathbf{\Sigma}^{1/2})} \geqslant \frac{\mu_{3i-2}(\mathbf{\Sigma}_{\mathbf{w}})}{\mu_i^2(\mathbf{\Sigma}^{1/2})} \geqslant \frac{\mu_{3i}(\mathbf{\Sigma}_{\mathbf{w}})}{\mu_i(\mathbf{\Sigma})}.$$

Combining the last two displays yields the desired result.

Proof of claim (15a)**.** Define the learning rate $\gamma_0(\Gamma) = \min\{1/[4\max_j \|\Gamma_{\cdot,j}\|_2^2], \gamma\}$ for any matrix $\Gamma \in \mathbb{R}^{M\times N}$ and define $\gamma_t(\Gamma)$ for all $t \in [L]$ according to (2). Let $\mathbf{\Sigma} = \mathbf{U}\tilde{\mathbf{\Gamma}}\mathbf{U}^\top$ be the singular value decomposition with $\mathbf{U}\mathbf{U}^\top = \mathbf{I}_M$ and $\tilde{\mathbf{\Gamma}} = \mathrm{diag}\{\tilde{\lambda}_1, \ldots, \tilde{\lambda}_M\}$ being a diagonal matrix with $\tilde{\lambda}_1 \geqslant \tilde{\lambda}_2 \geqslant \ldots \tilde{\lambda}_M \geqslant 0$. Note that $\mathbf{S}\mathbf{X}^\top/\sqrt{N} \overset{d}{=} \mathbf{U}\tilde{\mathbf{\Gamma}}^{1/2}\mathbf{Z}^\top$ conditioned on $\mathbf{S}$ and $\mathbf{U}^\top\widehat{\mathbf{\Sigma}}\mathbf{U} \overset{d}{=} \tilde{\mathbf{\Gamma}}^{1/2}\mathbf{Z}^\top\mathbf{Z}\tilde{\mathbf{\Gamma}}^{1/2}$. Therefore

$$\begin{aligned}
\mathbb{E}_{\mathbf{X}}[\mathsf{C}_L] &= \mathbb{E}_{\mathbf{X}}\big[\prod_{t=1}^{L}(\mathbf{I} - \gamma_t(\mathbf{S}\mathbf{X}^\top)\widehat{\mathbf{\Sigma}})\big] = \mathbf{U}\mathbb{E}_{\mathbf{X}}\big[\prod_{t=1}^{L}(\mathbf{I} - \gamma_t(\sqrt{N}\mathbf{U}\tilde{\mathbf{\Gamma}}^{1/2}\mathbf{Z}^\top)\mathbf{U}^\top\widehat{\mathbf{\Sigma}}\mathbf{U})\big]\mathbf{U}^\top \\
&= \mathbf{U}\mathbb{E}_{\mathbf{Z}}\big[\prod_{t=1}^{L}(\mathbf{I} - \gamma_t(\sqrt{N}\tilde{\mathbf{\Gamma}}^{1/2}\mathbf{Z}^\top)\tilde{\mathbf{\Gamma}}^{1/2}\mathbf{Z}^\top\mathbf{Z}\tilde{\mathbf{\Gamma}}^{1/2})\big]\mathbf{U}^\top.
\end{aligned}$$

Adopt the shorthand notation $\boldsymbol{\mathcal{U}} = \mathbf{Z}\tilde{\mathbf{\Gamma}}^{1/2}$ and write $\boldsymbol{\mathcal{U}} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N)^\top$. It suffices to show (note that γ_k is equal to γ_0 up to some k-dependent constant factor)

$$\mathbf{M} := \mathbb{E}_{\mathbf{Z}}[\gamma_0(\sqrt{N}\boldsymbol{\mathcal{U}}^\top)^K \cdot (\boldsymbol{\mathcal{U}}^\top\boldsymbol{\mathcal{U}})^K]$$

is a diagonal matrix for any $K \geqslant 0$. Consider the kl-entry $\mathbf{M}_{kl}$. It can be written as the sum of terms of the form $\mu_{i_1,j_1}\mu_{i_2,j_2}\cdots\mu_{i_{2K},j_{2K}}$ with $j_1 = k, j_{2K} = l, j_{2m} = j_{2m+1}, m \in [K-1]$. When $k \neq l$, there exists some $i \in [N]$ such that $\mu_{i,k}$ appears odd number of times in the product. Since flipping the sign of $\mu_{i,k}$ does not change $\gamma_0(\sqrt{N}\boldsymbol{\mathcal{U}}^\top)$, and $\mu_{i,j}$ are independent symmetric Gaussian variables, it follows that $\mathbb{E}_{\mathbf{Z}}[\gamma_0(\sqrt{N}\boldsymbol{\mathcal{U}}^\top)^K\mu_{i_1,j_1}\mu_{i_2,j_2}\cdots\mu_{i_{2K},j_{2K}}] = 0$. Consequently, we conclude that $\mathbf{M}_{kl} = 0$ for $k \neq l$ and $\mathbf{M}$ is a diagonal matrix.

Proof of claim (7b)**.** Let $\widehat{\mathbf{\Sigma}} = \widehat{\mathbf{U}}\widehat{\mathbf{\Gamma}}\widehat{\mathbf{U}}^\top$ be the singular value decomposition with $\widehat{\mathbf{\Gamma}} = \mathrm{diag}\{\widehat{\lambda}_1, \ldots, \widehat{\lambda}_M\}$ and $\widehat{\lambda}_1 \geqslant \ldots \geqslant \widehat{\lambda}_M$. Then we have

$$\prod_{t=1}^{L}(\mathbf{I} - \gamma_t\widehat{\mathbf{\Sigma}}) \succeq (\mathbf{I} - 2\gamma_0\widehat{\mathbf{\Sigma}})^{L_{\mathtt{eff}}} = \widehat{\mathbf{U}}(\mathbf{I} - 2\gamma_0\widehat{\mathbf{\Gamma}})^{L_{\mathtt{eff}}}\widehat{\mathbf{U}}^\top \succeq \frac{(\mathbf{I}_M - \widehat{\mathbf{U}}_{0:t}\widehat{\mathbf{U}}_{0:t}^\top)}{e},$$

where $\tilde{t} := \#\{i \in [M] : \widehat{\lambda}_i L_{\mathtt{eff}}\gamma_0 > 1/4\}$. Here, the first inequality follows from $\prod_{k=0}^{\lfloor \log L_{\mathtt{eff}} \rfloor - 1}(\mathbf{I} - \gamma_{kL_{\mathtt{eff}}+1}\widehat{\mathbf{\Sigma}}) \succeq \mathbf{I} - 2\gamma_0\widehat{\mathbf{\Sigma}}$ since $(1-t_1)(1-t_2) \geqslant 1 - t_1 - t_2$ for all $t_1, t_2 \in [0,1]$; the second inequality uses $(1-x)^{L_{\mathtt{eff}}} \geqslant \exp(-2L_{\mathtt{eff}}x) \geqslant e^{-1}$ for $x \in [0, 1/(2L_{\mathtt{eff}})]$. Therefore,

$$\mathbb{E}_{\mathbf{X}}[\mathsf{C}_L] = \mathbb{E}_{\mathbf{X}}(\mathbf{I} - \gamma_0\widehat{\mathbf{\Sigma}})^L \succeq \frac{1}{e}\mathbf{I}_M - \frac{1}{e}\mathbb{E}[\widehat{\mathbf{U}}_{0:t}\widehat{\mathbf{U}}_{0:t}^\top].$$

Since $\mathrm{tr}(\mathbb{E}[\widehat{\mathbf{U}}_{0:t}\widehat{\mathbf{U}}_{0:t}^\top]) = \mathbb{E}[\tilde{t}] = \bar{t}$, it follows that $\mathbb{E}[\widehat{\mathbf{U}}_{0:t}\widehat{\mathbf{U}}_{0:t}^\top]$ has at most $2\mathbb{E}[\tilde{t}]$ eigenvalues greater than $1/2$. Since $X \succeq Y \succeq \mathbf{0}_M$ implies $\mu_i(X) \geqslant \mu_i(Y)$ for all $i \in [M], X, Y \in \mathbb{R}^{M\times M}$ by Weyl's inequality, it follows that

$$\mu_{M-i+1}(\mathbb{E}_{\mathbf{X}}[\mathsf{C}_L]) \geqslant \frac{1}{2e}$$

for all $i \geqslant 2\mathbb{E}_{\mathbf{X}}[\#\{i \in [M] : \widehat{\lambda}_i L\gamma_0 > 1/4\}] + 1$.

□

B.3 Bias error under the source condition

Lemma B.3 (Bias bounds under the source condition)**.** *Let Assumption 1 hold, $a > b-1$, and assume $L_{\mathtt{eff}} \lesssim N^a/\gamma$. Under the notation in Theorem 3.1 and its proof, there exist some (a,b)-dependent constants $c, c' > 0$ such that when $\gamma \leqslant c/\log N$,*

$$\mathbb{E}_{\mathbf{w}^*}[\mathsf{Bias}(\mathbf{w}^*)] \lesssim \max\{(L_{\mathtt{eff}}\gamma)^{(1-b)/a}, M^{1-b}\},$$
$$\mathbb{E}_{\mathbf{w}^*}[\mathsf{Bias}(\mathbf{w}^*)] \gtrsim (L_{\mathtt{eff}}\gamma)^{(1-b)/a} \ \textit{when} \ (L_{\mathtt{eff}}\gamma)^{1/a} \leqslant M/c'.$$

with probability at least $1 - \exp(-\Omega(M))$ over the randomness of $\mathbf{S}$.

Proof of Lemma B.3. The proof follows from applying Lemma B.1 and Lemma B.2 under Assumption 1. We begin by verifying the conditions required in these two lemmas.

Verification of conditions (1)–(4) in Assumption 3. First, by Lemma E.5, we have $\mu_j(\mathbf{\Sigma}) \eqsim j^{-a}$ with probability at least $1 - \exp(-\Omega(M))$ over the randomness of $\mathbf{S}$. Since $a > 1$, it follows that $\gamma \lesssim 1 \lesssim \min\{1, c/\mathrm{tr}(\mathbf{\Sigma})\}$ and $\mathrm{tr}(\mathbf{\Sigma}^2) \lesssim 1$. Thus, conditions (1) and (2) in Assumption 3 are satisfied. Moreover, when $L \lesssim N^a$, we have

$$\begin{aligned}
&\sum_{i=1}^{M}\frac{\mu_i(\mathbf{\Sigma})}{\mu_i(\mathbf{\Sigma}) + 1/(L_{\mathtt{eff}}\gamma)} \\
&\leqslant \#\{i \in [M] : \mu_i(\mathbf{\Sigma}) \geqslant 1/(L_{\mathtt{eff}}\gamma)\} + (L_{\mathtt{eff}}\gamma)\cdot\sum_{i:\mu_i(\mathbf{\Sigma})<1/(L_{\mathtt{eff}}\gamma)}\mu_i(\mathbf{\Sigma}) \\
&\lesssim (L_{\mathtt{eff}}\gamma)^{1/a} + (L_{\mathtt{eff}}\gamma)\cdot\sum_{i:i\gtrsim(L_{\mathtt{eff}}\gamma)^{1/a}}\mu_i(\mathbf{\Sigma}) \lesssim (L_{\mathtt{eff}}\gamma)^{1/a} + (L_{\mathtt{eff}}\gamma)\cdot\sum_{i:i\gtrsim(L_{\mathtt{eff}}\gamma)^{1/a}} i^{-a} \\
&\lesssim (L_{\mathtt{eff}}\gamma)^{1/a} \leqslant N/4,
\end{aligned}$$

where the last inequality follows since we may assume $L_{\mathtt{eff}} \leqslant \tilde{c}N^a/\gamma$ for some constant $\tilde{c} > 0$ sufficiently small. Thus, condition (3) in Assumption 3 is satisfied.

To verify condition (4) in Assumption 3, we introduce $\mathbf{z}_1, \ldots \mathbf{z}_N \overset{iid}{\sim} \mathcal{N}(0, \mathbf{I}_M/N)$ and write $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_N)^\top$. It can be shown that $\mathbf{X}\mathbf{S}^\top/\sqrt{N} \overset{d}{=} \mathbf{Z}\mathbf{\Sigma}^{1/2}$ conditioned on $\mathbf{S}$. Therefore, we have $\mathrm{tr}(\widehat{\mathbf{\Sigma}}) \overset{d}{=} \mathrm{tr}(\mathbf{Z}\mathbf{\Sigma}\mathbf{Z}^\top)$, where $\mathbf{Z} \in \mathbb{R}^{N\times M}$ is a Gaussian sketching matrix. When $\mu_i(\mathbf{\Sigma}) \eqsim i^{-a}$ for all

$i \in [M]$ (which happens with probability at least $1 - \exp(-\Omega(M))$ over $\mathbf{S}$ by Lemma E.5), we have by Hansen-Wright inequality (see e.g., exercise 2.17 in Wainwright (2019)) and a union bound that

$$N \max_{i \in [N]} \mathbf{Z}_{i,\cdot} \mathbf{\Sigma} \mathbf{Z}_{i,\cdot}^\top \lesssim \mathrm{tr}(\mathbf{\Sigma}) + \|\mathbf{\Sigma}\|_2 \cdot \log(N/\delta) + \|\mathbf{\Sigma}\|_F \cdot \sqrt{\log(N/\delta)} \lesssim 1 + \log(N/\delta).$$

with probability at least $1-\delta$ over the randomness of $\bar{\mathbf{Z}}$. Thus, there exist some a-dependent constants $c, c' > 0$ such that when $\gamma \leqslant c/\log N$,

$$\mathbb{P}(\gamma_0 < \gamma/t) = \mathbb{P}(t/(4\gamma) < \max_i \|\mathbf{S}\mathbf{x}_i\|_2^2) = \mathbb{P}(t/(4\gamma) < N \max_i \mathbf{Z}_{i,\cdot} \mathbf{\Sigma} \mathbf{Z}_{i,\cdot}^\top) \leqslant N^{-c't}$$

for all $t \geqslant 1$. Therefore, condition (4) is also satisfied.

The upper bound. By Lemma E.5, we have $B_{\mathbf{B}}$ in Lemma B.1 satisfies

$$B_{\mathbf{B}} \leqslant c \cdot \big(1 + {L_{\mathtt{eff}}}^2 \cdot (N^{-2a} + N^{-2a} + N^{-2a} + N^{-2a})\big) \leqslant c \cdot (1 + N^{2a-2a}) \lesssim 1$$

with probability at least $1 - \exp(-\Omega(M))$, where the second inequality uses $L_{\mathtt{eff}} \lesssim N^a$. Moreover, we have by Lemma E.6 that $\frac{\mu_{M/2}(\mathbf{A}_k)}{\mu_M(\mathbf{A}_k)} \lesssim 1$ with probability at least $1 - \exp(-\Omega(M))$.

Now, choosing $k = \min\{M/3, (L_{\mathtt{eff}}\gamma)^{1/a}\}$ in Lemma B.1, using Assumption 1D, and taking expectation over $\mathbf{w}^*$ yields

$$\mathbb{E}_{\mathbf{w}^*}[\mathsf{Bias}(\mathbf{w}^*)] \lesssim \frac{\max\{k^{a-b+1}, 1\}}{L_{\mathtt{eff}}\gamma} + k^{1-b} \lesssim \max\{(L_{\mathtt{eff}}\gamma)^{(1-b)/a}, M^{1-b}\}$$

with probability at least $1 - \exp(-\Omega(M))$ over the randomness of $\mathbf{S}$.

The lower bound. By Lemma B.2, we have

$$\mathbb{E}_{\mathbf{w}^*} \mathsf{Bias}(\mathbf{w}^*) \gtrsim \sum_{i=2\bar{t}+1}^{M} \frac{\mu_{3i}(\mathbf{\Sigma}_{\mathbf{w}})}{\mu_i(\mathbf{\Sigma})},$$

with probability at least $1 - e^{-\Omega(M)}$, where $\bar{t} = \mathbb{E}_{\mathbf{X}}[\#\{i \in [M] : \widehat{\lambda}_i L_{\mathtt{eff}} \gamma_0 > 1/4\}]$ and $\{\widehat{\lambda}_i, i \in [M]\}$ are the eigenvalues of $\widehat{\mathbf{\Sigma}}$. Since $\widehat{\mathbf{\Sigma}} \overset{d}{=} \mathbf{Z}\mathbf{\Sigma}\mathbf{Z}^\top$ conditioned on $\mathbf{S}$ by Lemma E.3, when $\mu_i(\mathbf{\Sigma}) \eqsim i^{-a}$ for all $i \in [M]$ (which happens with probability at least $1 - \exp(-\Omega(M))$ over $\mathbf{S}$ by Lemma E.5), we have by combining Lemma E.4 and E.3 with $k = N/c$ that

$$\begin{aligned}
\widehat{\lambda}_{2j-1} &\overset{d}{=} \mu_{2j-1}(\mathbf{Z}\mathbf{\Sigma}\mathbf{Z}^\top) \\
&\leqslant \mu_j(\mathbf{Z}_{0:k}\mathbf{\Sigma}_{0:k}\mathbf{Z}_{0:k}^\top) + \mu_j(\mathbf{Z}_{k:\infty}\mathbf{\Sigma}_{k:\infty}\mathbf{Z}_{k:\infty}^\top) \\
&\lesssim \Big(1 + \sqrt{\frac{k + \log(1/\delta)}{N}}\Big) \cdot \mu_j(\mathbf{\Sigma}) + \Big(N^{-a} + \frac{N + \log(1/\delta)}{N^{a+1}} + \frac{\sqrt{N^{2-2a} + N^{1-2a}\log(1/\delta)}}{N}\Big) \\
&\lesssim \Big(1 + \sqrt{\frac{\log(1/\delta)}{N}}\Big) \cdot j^{-a} + N^{-a}\Big(1 + \frac{\log(1/\delta)}{N} + \sqrt{\frac{\log(1/\delta)}{N}}\Big) \qquad (16)
\end{aligned}$$

for all $j \leqslant k$ with probability at least $1 - \delta$ over the randomness of $\mathbf{Z}$. Therefore, it can be verified by a standard truncation argument that

$$\bar{t} = \mathbb{E}_{\mathbf{X}}[\#\{i \in [M] : \widehat{\lambda}_i L_{\mathtt{eff}} \gamma_0 > 1/4\}] \lesssim (L_{\mathtt{eff}}\gamma)^{1/a}.$$

Thus, when $(L_{\mathtt{eff}}\gamma_0)^{1/a} \leqslant M/c$ for some sufficiently large constant $c > 0$, we have

$$\mathbb{E}_{\mathbf{w}^*} \mathsf{Bias}(\mathbf{w}^*) \gtrsim \sum_{i=2\bar{t}+1}^{M} \frac{\mu_{3i}(\mathbf{\Sigma}_{\mathbf{w}})}{\mu_i(\mathbf{\Sigma})} \gtrsim \sum_{i=c'(L_{\mathtt{eff}}\gamma)^{1/a}}^{M} \frac{\mu_{3i}(\mathbf{\Sigma}_{\mathbf{w}})}{\mu_i(\mathbf{\Sigma})} \gtrsim \sum_{i=c'(L_{\mathtt{eff}}\gamma)^{1/a}}^{M} \frac{i^{-a-b}}{i^{-a}} \gtrsim (L_{\mathtt{eff}}\gamma)^{(1-b)/a}$$

with probability at least $1 - \exp(-\Omega(M))$ over the randomness of $\mathbf{S}$, where the third inequality uses Lemma E.5 (with $\mathbf{H}$ replaced by $\mathbf{H}\mathbf{H}^{\mathbf{w}}\mathbf{H}$).

□

C Variance error

C.1 Upper and lower bounds

Lemma C.1 (An upper bound on the GD variance term)**.** *Suppose Assumption 1A and 3 hold and $L_{\mathtt{eff}} \lesssim N^a/\gamma$. Under the notation in Theorem 3.1 and its proof, the variance term*

$$\mathsf{Var} := \mathbb{E}[\mathrm{tr}(\mathbf{X}\mathbf{S}^\top \mathbf{V}(\widehat{\mathbf{\Sigma}})\mathbf{\Sigma}\mathbf{V}(\widehat{\mathbf{\Sigma}})\mathbf{S}\mathbf{X}^\top)] \lesssim \frac{D^{\mathsf{U}}}{N}, \quad \textit{and} \quad \mathsf{Var} \gtrsim \frac{D^{\mathsf{L}}}{N},$$

where

$$D^{\mathsf{U}} := \mathbb{E}_{\mathbf{X}}\Big[\#\{i \in [M] : \widehat{\lambda}_i L_{\mathtt{eff}}\gamma_0 > 1/4\} + (L_{\mathtt{eff}}\gamma_0) \sum_{i:\widehat{\lambda}_i L_{\mathtt{eff}}\gamma_0 \leqslant 1/4} \widehat{\lambda}_i\Big],$$

$$D^{\mathsf{L}} := \mathbb{E}_{\mathbf{X}}\left[(L_{\mathtt{eff}}\gamma_0)^2 \cdot \sum_{i:\widehat{\lambda}_i L_{\mathtt{eff}}\gamma_0 \leqslant 1/4} \mu_i(\mathbf{\Sigma}) \cdot \mu_i(\widehat{\mathbf{\Sigma}}) + \frac{1}{5} \cdot \sum_{i:\widehat{\lambda}_i L_{\mathtt{eff}}\gamma_0 > 1/4} \frac{\mu_i(\mathbf{\Sigma})}{\mu_i(\widehat{\mathbf{\Sigma}})}\right],$$

and $(\widehat{\lambda}_i)_{i=1}^M$ are the eigenvalues of $\widehat{\mathbf{\Sigma}}$.

Proof of Lemma C.1. Note that

$$\mathbf{V}(\widehat{\mathbf{\Sigma}}) = \frac{1}{N}\sum_{t=1}^{L} \gamma_t \cdot \prod_{i=t+1}^{L}(\mathbf{I} - \gamma_i\widehat{\mathbf{\Sigma}}) = \frac{(\mathbf{I} - \prod_{t=1}^{L}(\mathbf{I} - \gamma_t\widehat{\mathbf{\Sigma}}))\widehat{\mathbf{\Sigma}}^{-1}}{N}.$$

Adopt the shorthand V_L for $\mathbf{I} - \prod_{t=1}^{L}(\mathbf{I} - \gamma_t\widehat{\mathbf{\Sigma}})$. Reorganizing the terms, we have

$$\mathsf{Var} = N \cdot \mathbb{E}[\mathrm{tr}(\mathbf{V}(\widehat{\mathbf{\Sigma}})\mathbf{\Sigma}\mathbf{V}(\widehat{\mathbf{\Sigma}})\widehat{\mathbf{\Sigma}})] = \frac{1}{N} \cdot \mathbb{E}_{\mathbf{X}}[\mathrm{tr}(\mathbf{\Sigma}\mathsf{V}_L\widehat{\mathbf{\Sigma}}^{-1}\mathsf{V}_L)].$$

Let $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_M$ be the eigenvalues of $\widehat{\mathbf{\Sigma}}$ in non-increasing order, and let $\lambda > 0$ be some value which will be given later. We now derive an upper bound and a lower bound for the variance Var.

An upper bound. Continuing the calculation, we further have

$$\begin{aligned}\mathrm{tr}(\mathbf{\Sigma}\mathsf{V}_L\widehat{\mathbf{\Sigma}}^{-1}\mathsf{V}_L) &= \mathrm{tr}(\mathsf{V}_L\widehat{\mathbf{\Sigma}}^{-1/2}(\widehat{\mathbf{\Sigma}} + \lambda\mathbf{I})^{1/2}[(\widehat{\mathbf{\Sigma}} + \lambda\mathbf{I})^{-1/2}\mathbf{\Sigma}(\widehat{\mathbf{\Sigma}} + \lambda\mathbf{I})^{-1/2}](\widehat{\mathbf{\Sigma}} + \lambda\mathbf{I})^{1/2}\widehat{\mathbf{\Sigma}}^{-1/2}\mathsf{V}_L) \\ &\leqslant \|\mathbf{\Sigma}^{1/2}(\widehat{\mathbf{\Sigma}} + \lambda\mathbf{I})^{-1/2}\|_2^2 \cdot [\mathrm{tr}(\mathsf{V}_L^2 + \lambda\widehat{\mathbf{\Sigma}}^{-1}\mathsf{V}_L^2)].\end{aligned}$$

Similar to the proof of claim (7b) in Lemma B.2, it can be verified that $\mathsf{V}_L \preceq \mathbf{I} - (\mathbf{I} - 2\gamma_0\widehat{\mathbf{\Sigma}})^{L_{\mathtt{eff}}}$ under the condition $\gamma_0 \leqslant 1/[4\max_i \|\mathbf{S}\mathbf{x}_i\|_2^2] \leqslant 1/[4\,\mathrm{tr}(\widehat{\mathbf{\Sigma}})]$ and stepsize assumption (2). Since $(1 - (1 - \gamma_0 x)^{L_{\mathtt{eff}}})^2 \leqslant \min\{(xL_{\mathtt{eff}}\gamma_0)^2, 1\}$ for $x \in [0, 1/(2\gamma_0)]$ by Bernoulli's inequality and $\sup_{[0,1/\gamma_0]}[(1 - (1 - \gamma_0 x)^{L_{\mathtt{eff}}})/x] = L_{\mathtt{eff}}\gamma_0$, it follows that

$$\begin{aligned}\mathrm{tr}(\mathsf{V}_L^2 + \lambda\widehat{\mathbf{\Sigma}}^{-1}\mathsf{V}_L^2) &\leqslant \sum_{i=1}^{M}\Big[((1 - (1 - 2\gamma_0\widehat{\lambda}_i)^{L_{\mathtt{eff}}})^2 + \frac{\lambda(1 - (1 - 2\gamma_0\widehat{\lambda}_i)^{L_{\mathtt{eff}}})^2}{\widehat{\lambda}_i}\Big] \\ &\lesssim \sum_{i=1}^{M}\Big[(1 + \lambda L_{\mathtt{eff}}\gamma_0) \cdot \mathbf{1}_{\{\widehat{\lambda}_i L_{\mathtt{eff}}\gamma_0 > 1/4\}} \\ &\qquad + (\lambda L_{\mathtt{eff}}\gamma_0(\widehat{\lambda}_i L_{\mathtt{eff}}\gamma_0) + (\widehat{\lambda}_i L_{\mathtt{eff}}\gamma_0)^2) \cdot \mathbf{1}_{\{\widehat{\lambda}_i L_{\mathtt{eff}}\gamma_0 \leqslant 1/4\}}\Big].\end{aligned}$$

Choosing $\lambda = 1/(L_{\mathtt{eff}}\gamma) \leqslant 1/(L_{\mathtt{eff}}\gamma_0)$ yields

$$\begin{aligned}&\mathrm{tr}(\mathsf{V}_L^2 + \lambda\widehat{\mathbf{\Sigma}}^{-1}\mathsf{V}_L^2) \\ &\lesssim \#\{i \in [M] : \widehat{\lambda}_i L_{\mathtt{eff}}\gamma_0 > 1/4\} + (L_{\mathtt{eff}}\gamma_0)^2 \sum_{i:\widehat{\lambda}_i L_{\mathtt{eff}}\gamma_0 \leqslant 1/4} \widehat{\lambda}_i^2 + (L_{\mathtt{eff}}\gamma_0) \sum_{i:\widehat{\lambda}_i L_{\mathtt{eff}}\gamma_0 \leqslant 1/4} \widehat{\lambda}_i =: \widetilde{D}^{\mathsf{U}} \\ &\lesssim \#\{i \in [M] : \widehat{\lambda}_i L_{\mathtt{eff}}\gamma_0 > 1/4\} + (L_{\mathtt{eff}}\gamma_0) \sum_{i:\widehat{\lambda}_i L_{\mathtt{eff}}\gamma_0 \leqslant 1/4} \widehat{\lambda}_i =: \widetilde{D}^{\mathsf{U}} \qquad (17)\end{aligned}$$

Applying Lemma E.1 and noting that $\mathrm{tr}(\mathsf{V}_L^2 + \lambda\widehat{\mathbf{\Sigma}}^{-1}\mathsf{V}_L^2) \lesssim N$ as V_L has at most N non-zero eigenvalues, we obtain

$$\mathbb{E}_{\mathbf{X}}[\mathrm{tr}(\mathbf{\Sigma}\mathsf{V}_L\widehat{\mathbf{\Sigma}}^{-1}\mathsf{V}_L)] \lesssim \mathbb{E}_{\mathbf{X}}\Big[\#\{i\in[M]:\widehat{\lambda}_i L_{\mathtt{eff}}\gamma_0 > 1/4\} + (L_{\mathtt{eff}}\gamma_0)\sum_{i:\widehat{\lambda}_i L_{\mathtt{eff}}\gamma_0\leqslant 1/4}\widehat{\lambda}_i\Big].$$

A lower bound. Similarly, by Von-Neumann's trace inequality, we have

$$\begin{aligned}\mathrm{tr}(\mathbf{\Sigma}\mathsf{V}_L\widehat{\mathbf{\Sigma}}^{-1}\mathsf{V}_L) &\geqslant \sum_{i=1}^{M}\mu_i(\mathbf{\Sigma})\mu_{M-i+1}(\mathsf{V}_L^2\widehat{\mathbf{\Sigma}}^{-1}) \overset{(i)}{\geqslant} \sum_{i=1}^{M}\mu_i(\mathbf{\Sigma})\frac{\mu_{2(M-i)+1}(\mathsf{V}_L^2\widehat{\mathbf{\Sigma}}^{-2})}{\mu_{M-i+1}(\widehat{\mathbf{\Sigma}}^{-1})}\\ &\overset{(ii)}{=} \sum_{i=1}^{M}\mu_i(\mathbf{\Sigma})\cdot\mu_i(\widehat{\mathbf{\Sigma}})\cdot\mu_{2(M-i)+1}(\mathsf{V}_L^2\widehat{\mathbf{\Sigma}}^{-2}),\end{aligned}$$

where step (i) uses $\mu_{i+j+1}(XY)\leqslant\mu_{i+1}(X)\mu_{j+1}(Y)$ for any matrices X, Y, and step (ii) uses the fact that $\mu_i(\widehat{\mathbf{\Sigma}}) = 1/\mu_{M-i+1}(\widehat{\mathbf{\Sigma}}^{-1})$.

Note that $\mathsf{V}_L \succeq \mathbf{I} - (\mathbf{I}-\gamma_0\widehat{\mathbf{\Sigma}})^{L_{\mathtt{eff}}}$. Since $f(x) := (1-(1-\gamma_0 x)^{L_{\mathtt{eff}}})^2/x^2$ is a decreasing function on $[0, 1/\gamma_0]$ and (1) $f(x)\geqslant(L_{\mathtt{eff}}\gamma_0)^2/4$ when $L_{\mathtt{eff}}\gamma_0 x\leqslant 1/4$; (2) $f(x)\geqslant 1/(5x^2)$ when $L_{\mathtt{eff}}\gamma_0 x\geqslant 1/4$, we have

$$\mathrm{tr}(\mathbf{\Sigma}\mathsf{V}_L\widehat{\mathbf{\Sigma}}^{-1}\mathsf{V}_L) \geqslant \frac{(L_{\mathtt{eff}}\gamma_0)^2}{4}\cdot\sum_{i:\widehat{\lambda}_i L_{\mathtt{eff}}\gamma_0\leqslant 1/4}\mu_i(\mathbf{\Sigma})\cdot\mu_i(\widehat{\mathbf{\Sigma}}) + \frac{1}{5}\cdot\sum_{i:\widehat{\lambda}_i L_{\mathtt{eff}}\gamma_0>1/4}\frac{\mu_i(\mathbf{\Sigma})}{\mu_i(\widehat{\mathbf{\Sigma}})}.$$

Taking expectation over $\mathbf{X}$ yields the desired result.

□

C.2 Variance error under the source condition

Lemma C.2 (Variance bounds under the source condition)**.** *Let Assumption 1 hold and assume $L_{\mathtt{eff}}\lesssim N^a/\gamma$. Under the notation in Theorem 3.1 and its proof, there exists some (a,b)-dependent constant $c>0$ such that when $\gamma\leqslant c/\log N$,*

$$\mathsf{Var}\eqsim\frac{\min\{M,(L_{\mathtt{eff}}\gamma)^{1/a}\}}{N}$$

with probability at least $1-\exp(-\Omega(M))$ over the randomness of $\mathbf{S}$.

Proof of Lemma C.2. Similar to the proof of Lemma B.3, we can verify that conditions (1)–(4) in Assumption 3 are satisfied with probability at least $1-\exp(-\Omega(M))$ over the randomness of $\mathbf{S}$. From the expression of D^{U}, it is straightforward to see that $D^{\mathsf{U}}\leqslant M$. Moreover, applying Lemma C.2, Eq. (16) in the proof of Lemma B.3 and a truncation argument, we can show that

$$\begin{aligned}D^{\mathsf{U}} &= \mathbb{E}_{\mathbf{X}}\Big[\#\{i\in[M]:\widehat{\lambda}_i L_{\mathtt{eff}}\gamma_0>1/4\} + (L_{\mathtt{eff}}\gamma_0)\sum_{i:\widehat{\lambda}_i L_{\mathtt{eff}}\gamma_0\leqslant 1/4}\widehat{\lambda}_i\Big]\\ &\lesssim (L_{\mathtt{eff}}\gamma)^{1/a} + \mathbb{E}_{\mathbf{X}}\Big[(L_{\mathtt{eff}}\gamma)\cdot\sum_{i:i\gtrsim(L_{\mathtt{eff}}\gamma)^{1/a}}\widehat{\lambda}_i\Big]\lesssim(L_{\mathtt{eff}}\gamma)^{1/a}\end{aligned}$$

with probability at least $1-\exp(-\Omega(M))$ over the randomness of $\mathbf{S}$. Thus, we have obtained $D^{\mathsf{U}}\lesssim\min\{M,(L_{\mathtt{eff}}\gamma)^{1/a}\}$.

For the lower bound, when $(L_{\mathtt{eff}}\gamma)^{1/a}\leqslant M/c$ for some sufficiently large constant $c>0$, conditioned on $\mathbf{S}$ such that $\mu_j(\mathbf{\Sigma})\eqsim j^{-a}$ for $j\in[M]$ (which holds with probability at least $1-e^{-\Omega(M)}$ by Lemma E.5), we have by Lemma E.8 that $\mu_j(\widehat{\mathbf{\Sigma}})\eqsim j^{-a}$ for $j\leqslant\min\{M,N\}/\tilde{c}$ with probability at least $1-e^{-\Omega(M)}$ for some $\tilde{c}>0$. Therefore,

$$D^{\mathsf{L}}\geqslant\mathbb{E}_{\mathbf{X}}\left[(L_{\mathtt{eff}}\gamma_0)^2\cdot\sum_{i:\widehat{\lambda}_i L_{\mathtt{eff}}\gamma_0\leqslant 1/4}\mu_i(\mathbf{\Sigma})\cdot\mu_i(\widehat{\mathbf{\Sigma}})\right]$$

$$\gtrsim (L_{\mathtt{eff}}\gamma)^2 \cdot \sum_{i:i\gtrsim (L_{\mathtt{eff}}\gamma)^{1/a}, i\leqslant \min\{M,N\}/\bar{c}} i^{-a}\cdot i^{-a} \gtrsim (L_{\mathtt{eff}}\gamma)^{1/a},$$

where the last line follows since we assume $(L_{\mathtt{eff}}\gamma)^{1/a} \leqslant M/c$ for some sufficiently large constant $c > 0$ and $(L_{\mathtt{eff}}\gamma)^{1/a} \lesssim N \lesssim N/c$.

On the other hand, similarly, when $(L_{\mathtt{eff}}\gamma)^{1/a} \geqslant M/c$ for some sufficiently constant $c > 0$, conditioned on $\mathbf{S}$ such that $\mu_j(\mathbf{\Sigma}) \eqsim j^{-a}$ for $j \in [M]$ (which holds with probability at least $1 - e^{-\Omega(M)}$ by Lemma E.5), we have by Lemma E.8 that $\mu_j(\widehat{\mathbf{\Sigma}}) \eqsim j^{-a}$ for $j \leqslant M$ with probability at least $1/2$. Therefore,

$$D^{\mathsf{L}} \geqslant \mathbb{E}_{\mathbf{X}}\left[\frac{1}{5}\cdot \sum_{i:\widehat{\lambda}_i L_{\mathtt{eff}}\gamma_0 > 1/4} \frac{\mu_i(\mathbf{\Sigma})}{\mu_i(\widehat{\mathbf{\Sigma}})}\right] \gtrsim \mathbb{E}_{\mathbf{X}}\left[\frac{1}{5}\cdot \sum_{i:i\leqslant M/c} \frac{\mu_i(\mathbf{\Sigma})}{\mu_i(\widehat{\mathbf{\Sigma}})}\right] \gtrsim M.$$

Putting pieces together yields the desired lower bound.

□

D Fluctuation error

D.1 An upper bound

Lemma D.1 (An upper bound on the fluctuation error)**.** *For each $i \in [N]$, define the leave-one-out GD process*

$$\boldsymbol{\theta}_t^{(-i)} = (\mathbf{I} - \gamma_t \widehat{\mathbf{\Sigma}}^{(-i)})\boldsymbol{\theta}_{t-1}^{(-i)} + \gamma_t (\mathbf{S}\mathbf{X}^\top \mathbf{y})^{(-i)}, \ \ \textit{with}\ \ \boldsymbol{\theta}_0^{(-i)} = \mathbf{0}, \tag{LOO-GD}$$

where $\widehat{\mathbf{\Sigma}}^{(-i)} := \sum_{j\neq i} \mathbf{S}\mathbf{x}_j\mathbf{x}_j^\top\mathbf{S}^\top/N$ and $(\mathbf{S}\mathbf{X}^\top\mathbf{y})^{(-i)} := \sum_{j\neq i}\mathbf{S}\mathbf{x}_j y_j/N$.

Let Assumption 1A, 1B, 3 hold and assume $L_{\mathtt{eff}} \lesssim N^a/\gamma$. Under the notation in Theorem 3.1 and its proof, for any $s \in [0,1], \alpha > 1$, there exists some (s,α)-dependent constant $c > 0$ such that the fluctuation error satisfies

$$\begin{aligned}\mathbb{E}[\mathsf{Fluc}] &= \mathbb{E}_{\mathbf{w}^*,(\mathbf{x}_i,y_i)_{i=1}^N, i_t, t\in[L]}[\|\mathbf{\Sigma}^{1/2}(\mathbf{v}_L - \boldsymbol{\theta}_L)\|_2^2] \\ &\leqslant c\cdot \mathbb{E}[\mathsf{F}\cdot \mathrm{tr}(\widehat{\mathbf{\Sigma}}^{1/\alpha})]\cdot \gamma^{1/\alpha} {L_{\mathtt{eff}}}^{1/\alpha - 1},\end{aligned}$$

with probability at least $1 - \exp(-\Omega(M))$ over the randomness of $\mathbf{S}$, where

$$\mathsf{F} := \widehat{R}\Big(\max_{i\in[N]}(\mathbf{x}_i^\top\mathbf{S}^\top\mathbf{v}^*)^2 + \max_{i\in[N]}\tilde{\epsilon}_i^2 + \max_{i\in[N],t\in[L]}(\mathbf{x}_i^\top\mathbf{S}^\top\boldsymbol{\theta}_t^{(-i)})^2 + \max_{i\in[N]}\|\mathbf{x}_i^\top\mathbf{S}^\top\|_{\mathbf{\Sigma}^{-s}}^2\cdot \mathbf{B}_{\mathbf{\Delta}}\Big),$$

$$\mathbf{B}_{\mathbf{\Delta}} := a_{\max}^2 \cdot \max_{i\in[N]}\|\mathbf{S}\mathbf{x}_i\|_2^2 \cdot \widehat{R}\cdot \frac{(L_{\mathtt{eff}}\gamma)^{2-s}}{N^2}, \ \ \textit{and}$$

$$a_{\max} := \max_{i\in[N],t\in[L]} |y_i + \mathbf{x}_i^\top\mathbf{S}^\top\boldsymbol{\theta}_t^{(-i)}|, \quad \lambda := \frac{1}{L_{\mathtt{eff}}\gamma}, \quad \widehat{R} := \|(\widehat{\mathbf{\Sigma}} + \lambda\mathbf{I})^{-1/2}(\mathbf{\Sigma} + \lambda\mathbf{I})^{1/2}\|_2^2.$$

Moreover, if $\mu_j(\widehat{\mathbf{\Sigma}}) \eqsim j^{-a}$ for $j \leqslant r(\widehat{\mathbf{\Sigma}})$ for some $a > 1$, then

$$\mathbb{E}_{i_t,t\in[L]}\|\mathbf{\Sigma}^{1/2}(\mathbf{v}_L - \boldsymbol{\theta}_L)\|_2^2 \leqslant c'\mathbb{E}[\mathsf{F}]\cdot\gamma^{1/a}{L_{\mathtt{eff}}}^{1/a-1}$$

for some a-dependent constant $c' > 0$.

Proof of Lemma D.1. The proof of this lemma follows from similar ideas as in the proof of Lemma 5 in Pillaud-Vivien et al. (2018), but with a more precise characterization on the magnitude of GD outputs. We start with an overview of the proof. At a high level, to bound the fluctuation error, we express the difference between the multi-pass SGD and GD trajectories, $\mathbf{v}_t - \boldsymbol{\theta}_t$, as a stochastic process (Eq. 18) that fits into the framework of Lemma D.2, which provides an upper bound on the fluctuation error $\mathbb{E}[\|\mathbf{\Sigma}^{1/2}(\mathbf{v}_t - \boldsymbol{\theta}_t)\|^2]$ under certain conditions, up to a mismatch between $\widehat{\mathbf{\Sigma}}$ and $\mathbf{\Sigma}$. We verify that the required conditions hold with appropriate choices of parameters (Eq. 20), which are further bounded using a leave-one-out argument (Lemma D.3). Applying Lemma D.2 with these parameters and a covariance replacement trick (Eq. 21) yields the desired bounds.

We now proceed to the proof. Define $\boldsymbol{\Delta}_t := \mathbf{v}_t - \boldsymbol{\theta}_t$ for $t \in [0, L]$. Recall that $\mathbf{v}^* = (\mathbf{SHS}^\top)^{-1}\mathbf{SHw}^*$ and we have $y_i = (\mathbf{S}\mathbf{x}_i)^\top \mathbf{v}^* + \tilde{\epsilon}_i$ for all $i \in [N]$ with $\tilde{\epsilon}_i$ independent of $\mathbf{S}\mathbf{x}_i$ conditioned on $(\mathbf{S}, \mathbf{w}^*)$ under the Gaussian assumption in Assumption 1A. Moreover, $\mathbb{E}[\tilde{\epsilon}_i|\mathbf{S}, \mathbf{w}^*] = 0$ and $\mathbb{E}[\tilde{\epsilon}_i^2|\mathbf{S}, \mathbf{w}^*] \leqslant \sigma^2 + \|\mathbf{w}^*\|_{\mathbf{H}}^2 =: \tilde{\sigma}^2(\mathbf{w}^*)$. By the definition of $\mathbf{v}_t$ and $\boldsymbol{\theta}_t$ in (multi-pass SGD) and (GD), we have

$$\boldsymbol{\Delta}_t = (\mathbf{I} - \gamma_t \mathbf{S}\mathbf{x}_{i_t}\mathbf{x}_{i_t}^\top \mathbf{S}^\top)\boldsymbol{\Delta}_{t-1} + \gamma_t \cdot (\boldsymbol{\xi}_{1,t} + \boldsymbol{\xi}_{2,t}), \tag{18}$$

where

$$\boldsymbol{\Delta}_0 = \mathbf{0}, \;\; \boldsymbol{\xi}_{1,t} := -\Big[\mathbf{S}\mathbf{x}_{i_t}\mathbf{x}_{i_t}^\top \mathbf{S}^\top - \widehat{\boldsymbol{\Sigma}}\Big](\boldsymbol{\theta}_{t-1} - \mathbf{v}^*), \quad \text{and} \quad \boldsymbol{\xi}_{2,t} := \mathbf{S}\mathbf{x}_{i_t}\tilde{\epsilon}_{i_t} - \mathbf{S}\mathbf{X}^\top \tilde{\boldsymbol{\epsilon}}/N, \;\; t \in [L].$$

Note that conditioned on $\mathbf{w}^*$, $\mathbf{S}$ and the dataset $\mathcal{D} = (\mathbf{x}_i, y_i)_{i=1}^N$, the noise terms $\mathbb{E}[\boldsymbol{\xi}_{1,t}|\mathbf{S}, \mathbf{w}^*, \mathcal{D}] = \mathbb{E}[\boldsymbol{\xi}_{2,t}|\mathbf{S}, \mathbf{w}^*, \mathcal{D}] = 0$. Next, we present the following two results.

Lemma D.2 (A modified Proposition 1 of Pillaud-Vivien et al. (2018) for the last iterate)**.** *Consider any recursion of the form*

$$\boldsymbol{\mu}_t = (\mathbf{I} - \gamma_t \cdot \boldsymbol{\nu}_t \boldsymbol{\nu}_t^\top)\boldsymbol{\mu}_{t-1} + \gamma_t \cdot \boldsymbol{\xi}_t, \quad \boldsymbol{\mu}_0 = \mathbf{0}, \quad t \in [L], \tag{19}$$

where the learning rates $(\gamma_t)_{t=1}^L$ are as defined in Theorem 3.1 and Eq. (2)*, $(\boldsymbol{\nu}_t, \boldsymbol{\xi}_t)_{t=1}^L \in \mathbb{R}^M \times \mathbb{R}^M$ are independent random vectors. Assume that $\mathbb{E}[\boldsymbol{\nu}_t \boldsymbol{\nu}_t^\top] = \boldsymbol{\Sigma}_{\boldsymbol{\nu}}$, $\mathbb{E}[\boldsymbol{\xi}_t] = \mathbf{0}$, $\mathbb{E}[\boldsymbol{\nu}_t \boldsymbol{\nu}_t^\top \boldsymbol{\nu}_t \boldsymbol{\nu}_t^\top] \preceq B_{\boldsymbol{\nu}}^2 \boldsymbol{\Sigma}_{\boldsymbol{\nu}}$, $\mathbb{E}[\boldsymbol{\xi}_t \boldsymbol{\xi}_t^\top] \preceq \sigma_\xi^2 \boldsymbol{\Sigma}_{\boldsymbol{\nu}}$, and $\gamma_0 B_{\boldsymbol{\nu}}^2 \leqslant 1/4$. Then for any $u \in [0, 1]$, we have*

$$\mathbb{E}[\|\boldsymbol{\Sigma}_{\boldsymbol{\nu}}^{u/2} \boldsymbol{\mu}_L\|_2^2] \leqslant c\sigma_\xi^2 \cdot \gamma \operatorname{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\nu}}^{1/\alpha})(L_{\mathtt{eff}}\gamma)^{1/\alpha - u}$$

fot any $\alpha > 1$ and some α-dependent constant $c > 0$. Moreover, there exists some a-dependent constant $c', \tilde{c} > 1$ such that when $\mu_j(\boldsymbol{\Sigma}_{\boldsymbol{\nu}}) \eqsim j^{-a}$ for $j \leqslant \min\{M, N/\tilde{c}\}$, we have

$$\mathbb{E}[\|\boldsymbol{\Sigma}_{\boldsymbol{\nu}}^{u/2} \boldsymbol{\mu}_L\|_2^2] \leqslant c'\sigma_\xi^2 \cdot \gamma (L_{\mathtt{eff}}\gamma)^{1/a - u}$$

for any $u \in [0, 1]$ and some a-dependent constant $c' > 0$.

See the proof of Lemma D.2 in Section D.4.

Lemma D.3 (A leave-one-out bound on GD iterates)**.** *Under the assumptions and notation in Lemma D.1, for any $s \in [0, 1]$, there exists some s-dependent constant $c > 0$ such that the* (GD) *updates $(\boldsymbol{\theta}_t)_{t=1}^L$ satisfies*

$$\max_{i \in [N], t \in [L]} (\mathbf{x}_i^\top \mathbf{S}^\top \boldsymbol{\theta}_t)^2 \leqslant c \cdot \Big[\max_{i \in [N], t \in [L]} (\mathbf{x}_i^\top \mathbf{S}^\top \boldsymbol{\theta}_t^{(-i)})^2 + \max_{i \in [N]} \|\mathbf{x}_i^\top \mathbf{S}^\top\|_{\boldsymbol{\Sigma}^{-s}}^2 \cdot \mathbf{B}_{\boldsymbol{\Delta}}\Big].$$

See the proof of Lemma D.3 in Section D.5.

Let $\boldsymbol{\nu}_t = \mathbf{S}\mathbf{x}_{i_t}$, $\boldsymbol{\xi}_t = \boldsymbol{\xi}_{1,t} + \boldsymbol{\xi}_{2,t}$. We claim that $(\boldsymbol{\nu}_t, \boldsymbol{\xi}_t)$ satisfies the conditions in Lemma D.2 with

$$\boldsymbol{\Sigma}_{\boldsymbol{\nu}} = \widehat{\boldsymbol{\Sigma}}, \;\; B_{\boldsymbol{\nu}} = \max_{i \in [N]} \|\mathbf{S}\mathbf{x}_i\|_2, \;\; \sigma_\xi^2 = 2 \max_{i \in [N], t \in [L]} [(\mathbf{x}_i^\top \mathbf{S}^\top (\boldsymbol{\theta}_{t-1} - \mathbf{v}^*))^2 + \tilde{\epsilon}_i^2]. \tag{20}$$

Thus applying Lemma D.2 with $u = 0, 1$ to the stochastic process in (18) and letting $\lambda = \frac{1}{L_{\mathtt{eff}}\gamma}$ yields

$$\begin{aligned}\mathbb{E}_{i_t, t \in [L]} \|\boldsymbol{\Sigma}^{1/2}(\mathbf{v}_L - \boldsymbol{\theta}_L)\|_2^2 &\lesssim \|(\widehat{\boldsymbol{\Sigma}} + \lambda \mathbf{I})^{-1/2}(\boldsymbol{\Sigma} + \lambda \mathbf{I})^{1/2}\|^2 \cdot \mathbb{E}_{i_t, t \in [L]} \|(\widehat{\boldsymbol{\Sigma}} + \lambda \mathbf{I})^{1/2}(\mathbf{v}_L - \boldsymbol{\theta}_L)\|_2^2 \\ &\lesssim \sigma_\xi^2 \cdot \|(\widehat{\boldsymbol{\Sigma}} + \lambda \mathbf{I})^{-1/2}(\boldsymbol{\Sigma} + \lambda \mathbf{I})^{1/2}\|^2 \cdot \gamma \operatorname{tr}(\widehat{\boldsymbol{\Sigma}}^{1/\alpha})(L_{\mathtt{eff}}\gamma)^{1/\alpha - 1}.\end{aligned} \tag{21}$$

Moreover,

$$\begin{aligned}\sigma_\xi^2 &\lesssim \max_{i \in [N]} (\mathbf{x}_i^\top \mathbf{S}^\top \mathbf{v}^*)^2 + \max_{i \in [N]} \tilde{\epsilon}_i^2 + \max_{i \in [N], t \in [L]} (\mathbf{x}_i^\top \mathbf{S}^\top \boldsymbol{\theta}_t)^2 \\ &\lesssim \max_{i \in [N]} (\mathbf{x}_i^\top \mathbf{S}^\top \mathbf{v}^*)^2 + \max_{i \in [N]} \tilde{\epsilon}_i^2 + \max_{i \in [N], t \in [L]} (\mathbf{x}_i^\top \mathbf{S}^\top \boldsymbol{\theta}_t^{(-i)})^2 + \max_{i \in [N]} \|\mathbf{x}_i^\top \mathbf{S}^\top\|_{\boldsymbol{\Sigma}^{-s}}^2 \cdot \mathbf{B}_{\boldsymbol{\Delta}},\end{aligned}$$

where the second line follows from Lemma D.3. Putting the last two displays together and taking expectation over $(\mathbf{x}_i, y_i)_{i=1}^N, \mathbf{w}^*$ yields the first part of Lemma D.1. The second part of Lemma D.1 follows from the same argument by applying the second part of Lemma D.2.

Proof of claim (20)**.** Conditioned on $\mathbf{S}$ and $(\mathbf{x}_i, y_i)_{i=1}^N$, when choosing $\boldsymbol{\nu}_t = \mathbf{S}\mathbf{x}_{i_t}$, we have $\mathbb{E}[\boldsymbol{\nu}_t\boldsymbol{\nu}_t^\top] = \mathbb{E}[\mathbf{S}\mathbf{x}_{i_t}\mathbf{x}_{i_t}^\top\mathbf{S}^\top] = \widehat{\boldsymbol{\Sigma}}$, and $\mathbb{E}[\boldsymbol{\nu}_t\boldsymbol{\nu}_t^\top\boldsymbol{\nu}_t\boldsymbol{\nu}_t^\top] \leqslant \mathbb{E}[\max_{i\in[N]}\|\boldsymbol{\nu}_i\|^2\boldsymbol{\nu}_i\boldsymbol{\nu}_i^\top] \preceq \max_{i\in[N]}\|\mathbf{S}\mathbf{x}_i\|_2^2 \cdot \widehat{\boldsymbol{\Sigma}} = B_{\boldsymbol{\nu}}^2 \cdot \widehat{\boldsymbol{\Sigma}}$. Thus we may let $\boldsymbol{\Sigma}_{\boldsymbol{\nu}} = \widehat{\boldsymbol{\Sigma}}$ and $B_{\boldsymbol{\nu}} = \max_{i\in[N]}\|\mathbf{S}\mathbf{x}_i\|_2$. In this case, we have $\gamma_0 B_{\boldsymbol{\nu}}^2 \leqslant 1/4$ by the assumption that $\gamma_0 \leqslant 1/[4\max_{i\in[N]}\|\mathbf{S}\mathbf{x}_i\|_2^2]$. It remains to bound σ_ξ^2 in (20). Note that

$$\begin{aligned}
\mathbb{E}[\boldsymbol{\xi}_t\boldsymbol{\xi}_t^\top] &\leq 2\mathbb{E}[\boldsymbol{\xi}_{1,t}\boldsymbol{\xi}_{1,t}^\top] + 2\mathbb{E}[\boldsymbol{\xi}_{2,t}\boldsymbol{\xi}_{2,t}^\top] \\
&\leq 2\mathbb{E}[\boldsymbol{\nu}_i\boldsymbol{\nu}_i^\top(\boldsymbol{\theta}_{t-1}-\mathbf{v}^*)(\boldsymbol{\theta}_{t-1}-\mathbf{v}^*)^\top\boldsymbol{\nu}_i\boldsymbol{\nu}_i^\top] + 2\mathbb{E}[\mathbf{S}\mathbf{x}_{i_t}^\top\tilde{\boldsymbol{\epsilon}}_{i_t}(\mathbf{S}\mathbf{x}_{i_t}^\top\tilde{\boldsymbol{\epsilon}}_{i_t})^\top] \\
&\leq 2\max_{i\in[N]}(\mathbf{x}_i^\top\mathbf{S}^\top(\boldsymbol{\theta}_{t-1}-\mathbf{v}^*))^2 \cdot \boldsymbol{\Sigma}_{\boldsymbol{\nu}} + 2\max_{i\in[N]}\tilde{\epsilon}_i^2 \cdot \boldsymbol{\Sigma}_{\boldsymbol{\nu}},
\end{aligned}$$

where the second line uses Jensen's inequality. Therefore, we can set

$$\sigma_\xi^2 = 2\max_{i\in[N],t\in[L]}[(\mathbf{x}_i^\top\mathbf{S}^\top(\boldsymbol{\theta}_{t-1}-\mathbf{v}^*))^2 + \tilde{\epsilon}_i^2]$$

and the conditions required by Lemma D.2 are satisfied.

□

D.2 A lower bound

Lemma D.4 (A lower bound on the fluctuation error)**.** *Let Assumption 1A, 1B, 3 hold and assume* $L_{\mathtt{eff}} \lesssim N^a/\gamma$*. Under the notation in Theorem 3.1 and its proof, with probability at least* $1 - \exp(-\Omega(M))$ *over the randomness of* $\mathbf{S}$*,*

$$\begin{aligned}
\mathbb{E}_{(\mathbf{x}_i,y_i)_{i\in[N]},i_t,t\in[L]}[\mathsf{Fluc}] &= \mathbb{E}_{(\mathbf{x}_i,y_i)_{i\in[N]},i_t,t\in[L]}[\|\boldsymbol{\Sigma}^{1/2}(\mathbf{v}_L-\boldsymbol{\theta}_L)\|_2^2] \\
&\gtrsim (\sigma^2 + \|\mathbf{w}^*\|_{\mathbf{H}}^2)\cdot\mathbb{E}_{(\mathbf{x}_i)_{i\in[N]}}\Big[\frac{B_{\boldsymbol{\xi}}\gamma_0 L_{\mathtt{eff}}\gamma_0}{10}\cdot\sum_{i>\tilde{t}}\mu_i(\widehat{\boldsymbol{\Sigma}})\cdot\mu_i(\boldsymbol{\Sigma})\Big],
\end{aligned}$$

where $B_{\boldsymbol{\xi}} := \max\{\frac{N-1}{N} - \frac{4\gamma_0 L_{\mathtt{eff}}}{N}B_{\boldsymbol{\nu}}^2, 0\}$*,* $B_{\boldsymbol{\nu}} := \max_{i\in[N]}\|\mathbf{S}\mathbf{x}_i\|_2$ *and* $\tilde{t} := \#\{i \in [M] : \widehat{\lambda}_i L_{\mathtt{eff}}\gamma_0 > 1/8\}$*, and* $(\widehat{\lambda}_i)_{i=1}^M$ *are the eigenvalues of* $\widehat{\boldsymbol{\Sigma}}$*.*

Proof of Lemma D.4. Define $\boldsymbol{\Delta}_t := \mathbf{v}_t - \boldsymbol{\theta}_t$ for $t \in [L]$. Similar to the proof of Lemma D.1, conditioned on $\mathbf{S}$ and $\mathbf{w}^*$, we have

$$\begin{aligned}
\boldsymbol{\Delta}_t &= (\mathbf{I} - \gamma_t\mathbf{S}\mathbf{x}_{i_t}\mathbf{x}_{i_t}^\top\mathbf{S}^\top)\boldsymbol{\Delta}_{t-1} + \gamma_t\cdot(\boldsymbol{\xi}_{1,t}+\boldsymbol{\xi}_{2,t}) \\
&= \sum_{i=1}^L \gamma_i \cdot \sum_{j=i+1}^L (\mathbf{I} - \gamma_j\mathbf{S}\mathbf{x}_{i_j}\mathbf{x}_{i_j}^\top\mathbf{S}^\top)^\top(\boldsymbol{\xi}_{1,i}+\boldsymbol{\xi}_{2,i})
\end{aligned} \tag{22}$$

where

$$\boldsymbol{\Delta}_0 = \mathbf{0},\ \ \boldsymbol{\xi}_{1,t} := -\Big[\mathbf{S}\mathbf{x}_{i_t}\mathbf{x}_{i_t}^\top\mathbf{S}^\top - \widehat{\boldsymbol{\Sigma}}\Big](\boldsymbol{\theta}_{t-1}-\mathbf{v}^*),\quad\text{and}\quad \boldsymbol{\xi}_{2,t} := \mathbf{S}\mathbf{x}_{i_t}\tilde{\epsilon}_{i_t} - \mathbf{S}\mathbf{X}^\top\tilde{\boldsymbol{\epsilon}}/N,\ \ t\in[L],$$

and $\tilde{\boldsymbol{\epsilon}}_i$ are i.i.d $\mathcal{N}(0, \tilde{\sigma}^2(\mathbf{w}^*))$ independent of $\mathbf{S}\mathbf{x}_i$ conditioned on $\mathbf{S}$ and $\mathbf{w}^*$, where $\tilde{\sigma}^2(\mathbf{w}^*) := \sigma^2 + \|\mathbf{w}^*\|_{\mathbf{H}}^2$. Let $B_{\boldsymbol{\nu}} := \max_{i\in[N]}\|\mathbf{S}\mathbf{x}_i\|_2$. We claim that

$$\mathbb{E}_{(\tilde{\boldsymbol{\epsilon}}_i)_{i\in[N]}}\mathbb{E}_{i_t,t\in[L]}[(\boldsymbol{\xi}_{1,i}+\boldsymbol{\xi}_{2,i})(\boldsymbol{\xi}_{1,i}+\boldsymbol{\xi}_{2,i})^\top] \succeq \tilde{\sigma}^2(\mathbf{w}^*)B_{\boldsymbol{\xi}}\cdot\widehat{\boldsymbol{\Sigma}}, \tag{23}$$

and we have $B_{\boldsymbol{\xi}} \geqslant 1/2$ when $(L_{\mathtt{eff}}\gamma)B_{\boldsymbol{\nu}}^2/N \leqslant 1/3$. The proof of this claim is deferred to the end of the proof.

Since $\boldsymbol{\xi}_{1,t}$ and $\boldsymbol{\xi}_{2,t}$ are zero-mean noise, conditioned on $\mathbf{S}, \mathbf{w}^*$ and $(\mathbf{x}_i, y_i)_{i=1}^N$, we have

$$\begin{aligned}
&\mathbb{E}_{i_t,t\in[L]}\mathbb{E}_{(\tilde{\boldsymbol{\epsilon}}_k)_{k\in[N]}}[\|\boldsymbol{\Sigma}^{1/2}(\mathbf{v}_L-\boldsymbol{\theta}_L)\|_2^2] \\
&= \sum_{i=1}^L \gamma_i^2\cdot\mathbb{E}_{(\tilde{\boldsymbol{\epsilon}}_k)_{k\in[N]}}\mathbb{E}_{i_t,t\in[L]}[\mathrm{tr}(\prod_{j=i+1}^L(\mathbf{I}-\gamma_j\mathbf{S}\mathbf{x}_{i_j}\mathbf{x}_{i_j}^\top\mathbf{S}^\top)\boldsymbol{\Sigma}\prod_{j=i+1}^L(\mathbf{I}-\gamma_j\mathbf{S}\mathbf{x}_{i_j}\mathbf{x}_{i_j}^\top\mathbf{S}^\top)^\top(\boldsymbol{\xi}_{1,i}+\boldsymbol{\xi}_{2,i})(\boldsymbol{\xi}_{1,i}+\boldsymbol{\xi}_{2,i})^\top)]
\end{aligned}$$

$$\geq \tilde{\sigma}^2(\mathbf{w}^*)B_{\boldsymbol{\xi}} \cdot \sum_{i=1}^{L} \gamma_i^2 \mathbb{E}_{i_t, t\in[L]}[\mathrm{tr}(\prod_{j=i+1}^{L}(\mathbf{I}-\gamma_j \mathbf{S}\mathbf{x}_{i_j}\mathbf{x}_{i_j}^\top \mathbf{S}^\top)\boldsymbol{\Sigma}\prod_{j=i+1}^{L}(\mathbf{I}-\gamma_j \mathbf{S}\mathbf{x}_{i_j}\mathbf{x}_{i_j}^\top \mathbf{S}^\top)^\top \widehat{\boldsymbol{\Sigma}})]$$
$$\geq \tilde{\sigma}^2(\mathbf{w}^*)B_{\boldsymbol{\xi}} \cdot \sum_{i=1}^{L} \gamma_i^2 \,\mathrm{tr}(\boldsymbol{\Sigma}\widehat{\boldsymbol{\Sigma}}\prod_{j=i+1}^{L}(\mathbf{I}-2\gamma_j\widehat{\boldsymbol{\Sigma}})),$$

where the last line follows from the fact that $\mathbb{E}_{i_j}[(\mathbf{I}-\gamma_j \mathbf{S}\mathbf{x}_{i_j}\mathbf{x}_{i_j}^\top \mathbf{S}^\top)P(\widehat{\boldsymbol{\Sigma}})(\mathbf{I}-\gamma_j \mathbf{S}\mathbf{x}_{i_j}\mathbf{x}_{i_j}^\top \mathbf{S}^\top)] \succeq P(\widehat{\boldsymbol{\Sigma}})(\mathbf{I}-2\gamma_j\widehat{\boldsymbol{\Sigma}})$ for any polynomial P. Continuing from the last line, we have

$$\begin{aligned}
&\mathbb{E}_{i_t,t\in[L]}\mathbb{E}_{(\tilde{\boldsymbol{\epsilon}}_k)_{k\in[N]}}[\|\boldsymbol{\Sigma}^{1/2}(\mathbf{v}_L-\boldsymbol{\theta}_L)\|_2^2] \\
&\geq \tilde{\sigma}^2(\mathbf{w}^*)B_{\boldsymbol{\xi}} \cdot \mathrm{tr}\Big(\widehat{\boldsymbol{\Sigma}}\sum_{i=1}^{L}\gamma_i^2\prod_{j=i+1}^{L}(\mathbf{I}-2\gamma_j\widehat{\boldsymbol{\Sigma}})\boldsymbol{\Sigma}\Big) \\
&= \tilde{\sigma}^2(\mathbf{w}^*)B_{\boldsymbol{\xi}} \cdot \sum_{k=0}^{\lfloor \log L-1\rfloor}\gamma_{L_{\mathtt{eff}}k+1}^2 \cdot \mathrm{tr}\Big(\widehat{\boldsymbol{\Sigma}}\frac{\mathbf{I}-(\mathbf{I}-2\gamma_{L_{\mathtt{eff}}k+1}\widehat{\boldsymbol{\Sigma}})^{L_{\mathtt{eff}}}}{2\gamma_{L_{\mathtt{eff}}k+1}\widehat{\boldsymbol{\Sigma}}}\prod_{j=k+1}^{\lfloor \log L-1\rfloor}(\mathbf{I}-2\gamma_{L_{\mathtt{eff}}j+1}\widehat{\boldsymbol{\Sigma}})^{L_{\mathtt{eff}}}\,\boldsymbol{\Sigma}\Big) \\
&\geq \tilde{\sigma}^2(\mathbf{w}^*)B_{\boldsymbol{\xi}} \cdot \gamma_0\,\mathrm{tr}\Big((\mathbf{I}-(\mathbf{I}-2\gamma_0\widehat{\boldsymbol{\Sigma}})^{L_{\mathtt{eff}}})(\mathbf{I}-4\gamma_0\widehat{\boldsymbol{\Sigma}})^{L_{\mathtt{eff}}}\,\boldsymbol{\Sigma}\Big),
\end{aligned}$$

where the last line uses $\prod_{j=0}^{\lfloor \log L-1\rfloor}(1-2\gamma_{L_{\mathtt{eff}}j+1}x)^{L_{\mathtt{eff}}} \geqslant (1-4\gamma_0 x)^{L_{\mathtt{eff}}}$ for $x\in[0,1/(2\gamma_0)]$ by the stepsize definition (2). Since $\mu_{i+j+1}(XY)\leqslant \mu_{i+1}(X)\mu_{j+1}(Y)$ for any i,j and matrices X,Y of matching dimensions, it follows that for any $j\geqslant 1$

$$\begin{aligned}
&\mu_j((\mathbf{I}-(\mathbf{I}-2\gamma_0\widehat{\boldsymbol{\Sigma}})^{L_{\mathtt{eff}}})(\mathbf{I}-4\gamma_0\widehat{\boldsymbol{\Sigma}})^{L_{\mathtt{eff}}}\,\boldsymbol{\Sigma}) \\
&\geqslant \frac{\mu_{3j-2}((\mathbf{I}-(\mathbf{I}-2\gamma_0\widehat{\boldsymbol{\Sigma}})^{L_{\mathtt{eff}}})(\mathbf{I}-4\gamma_0\widehat{\boldsymbol{\Sigma}})^{L_{\mathtt{eff}}}/\widehat{\boldsymbol{\Sigma}})}{\mu_j(\widehat{\boldsymbol{\Sigma}}^{-1})\mu_j(\boldsymbol{\Sigma}^{-1})} \\
&= \mu_{M-j}(\widehat{\boldsymbol{\Sigma}})\cdot\mu_{M-j}(\boldsymbol{\Sigma})\cdot\mu_{3j-2}\Big(\frac{(\mathbf{I}-(\mathbf{I}-2\gamma_0\widehat{\boldsymbol{\Sigma}})^{L_{\mathtt{eff}}})(\mathbf{I}-4\gamma_0\widehat{\boldsymbol{\Sigma}})^{L_{\mathtt{eff}}}}{\widehat{\boldsymbol{\Sigma}}}\Big).
\end{aligned}$$

Since $f(x)=(1-(1-2\gamma_0 x)^{L_{\mathtt{eff}}})(1-4\gamma_0 x)^{L_{\mathtt{eff}}}/x$ satisfies $f(x)\geqslant L_{\mathtt{eff}}\gamma_0/10$ for $x\in[0,1/(8\gamma_0 L_{\mathtt{eff}})]$, it follows that $\frac{(\mathbf{I}-(\mathbf{I}-2\gamma_0\widehat{\boldsymbol{\Sigma}})^{L_{\mathtt{eff}}})(\mathbf{I}-4\gamma_0\widehat{\boldsymbol{\Sigma}})^{L_{\mathtt{eff}}}}{\widehat{\boldsymbol{\Sigma}}}$ has at most $\tilde{t}=\#\{i\in[M]:\widehat{\lambda}_i L_{\mathtt{eff}}\gamma_0>1/8\}$ eigenvalues that are less than $L_{\mathtt{eff}}\gamma_0/10$. Therefore, we have

$$\begin{aligned}
\mathrm{tr}\Big((\mathbf{I}-(\mathbf{I}-2\gamma_0\widehat{\boldsymbol{\Sigma}})^{L_{\mathtt{eff}}})(\mathbf{I}-4\gamma_0\widehat{\boldsymbol{\Sigma}})^{L_{\mathtt{eff}}}\,\boldsymbol{\Sigma}\Big) &= \sum_{j=1}^{M}\mu_j\Big((\mathbf{I}-(\mathbf{I}-2\gamma_0\widehat{\boldsymbol{\Sigma}})^{L_{\mathtt{eff}}})(\mathbf{I}-4\gamma_0\widehat{\boldsymbol{\Sigma}})^{L_{\mathtt{eff}}}\,\boldsymbol{\Sigma}\Big) \\
&\geqslant \frac{L_{\mathtt{eff}}\gamma_0}{10}\cdot\sum_{i>\tilde{t}}\mu_i(\widehat{\boldsymbol{\Sigma}})\cdot\mu_i(\boldsymbol{\Sigma}).
\end{aligned}$$

Putting pieces together and taking expectation over $(\mathbf{x}_i)_{i\in[N]}$, we obtain

$$\mathbb{E}_{(\mathbf{x}_i)_{i\in[N]}}\mathbb{E}_{(\tilde{\boldsymbol{\epsilon}}_k)_{k\in[N]}}[\|\boldsymbol{\Sigma}^{1/2}(\mathbf{v}_L-\boldsymbol{\theta}_L)\|_2^2] \gtrsim \tilde{\sigma}^2(\mathbf{w}^*)\cdot\mathbb{E}_{(\mathbf{x}_i)_{i\in[N]}}\Big[\frac{B_{\boldsymbol{\xi}}\gamma_0 L_{\mathtt{eff}}\gamma_0}{10}\cdot\sum_{i>\tilde{t}}\mu_i(\widehat{\boldsymbol{\Sigma}})\cdot\mu_i(\boldsymbol{\Sigma})\Big].$$

Proof of claim (23)**.** By Eq. (4) in the proof of Theorem 3.1, we have

$$\boldsymbol{\theta}_t-\mathbf{v}=-\prod_{i=1}^{t}\Big(\mathbf{I}-\gamma_i\widehat{\boldsymbol{\Sigma}}\Big)\mathbf{v}^*+\mathbf{V}_t(\widehat{\boldsymbol{\Sigma}})\mathbf{S}\mathbf{X}^\top\tilde{\boldsymbol{\epsilon}},$$

where

$$\mathbf{V}_t(\widehat{\boldsymbol{\Sigma}}):=\frac{1}{N}\sum_{i=1}^{t}\gamma_i\cdot\prod_{j=i+1}^{t}(\mathbf{I}-\gamma_j\widehat{\boldsymbol{\Sigma}})=\frac{\mathbf{I}-\prod_{i=1}^{t}(\mathbf{I}-\gamma_i\widehat{\boldsymbol{\Sigma}})}{N\widehat{\boldsymbol{\Sigma}}}.$$

Let

$$\boldsymbol{\xi}_i^s:=-(\mathbf{S}\mathbf{x}_{i_t}\mathbf{x}_{i_t}^\top\mathbf{S}^\top-\widehat{\boldsymbol{\Sigma}})\mathbf{V}_{t-1}(\widehat{\boldsymbol{\Sigma}})\mathbf{S}\mathbf{X}^\top\tilde{\boldsymbol{\epsilon}}+(\mathbf{S}\mathbf{x}_{i_t}\tilde{\epsilon}_{i_t}-\mathbf{S}\mathbf{X}^\top\tilde{\boldsymbol{\epsilon}}/N)$$

$$= \boldsymbol{\xi}_{1,i} + \boldsymbol{\xi}_{2,i} + (\mathbf{S}\mathbf{x}_{i_t}\mathbf{x}_{i_t}^\top\mathbf{S}^\top - \widehat{\boldsymbol{\Sigma}})\prod_{i=1}^{t}(\mathbf{I} - \gamma_i\widehat{\boldsymbol{\Sigma}})\mathbf{v}^*.$$

Since $\mathbb{E}_{i_t,t\in[L]}[\boldsymbol{\xi}_i^s] = 0$, it can be verified that

$$\begin{aligned}
&\mathbb{E}_{\tilde{\boldsymbol{\epsilon}}}\mathbb{E}_{i_t,t\in[L]}[(\boldsymbol{\xi}_{1,i} + \boldsymbol{\xi}_{2,i})(\boldsymbol{\xi}_{1,i} + \boldsymbol{\xi}_{2,i})^\top] \\
\geqslant\ &\mathbb{E}_{\tilde{\boldsymbol{\epsilon}}}\mathbb{E}_{i_t,t\in[L]}[\boldsymbol{\xi}_i^s\boldsymbol{\xi}_i^{s\top}] \\
\geq\ &\tilde{\sigma}^2(\mathbf{w}^*)\cdot\Big[\frac{N-1}{N}\widehat{\boldsymbol{\Sigma}} - (\mathbf{S}\mathbf{x}_{i_t}\mathbf{x}_{i_t}^\top\mathbf{S}^\top - \widehat{\boldsymbol{\Sigma}})\Big[\frac{\mathbf{I} - \prod_{i=1}^{t}(\mathbf{I}-\gamma_i\widehat{\boldsymbol{\Sigma}})^2}{N\widehat{\boldsymbol{\Sigma}}}\Big](\mathbf{S}\mathbf{x}_{i_t}\mathbf{x}_{i_t}^\top\mathbf{S}^\top - \widehat{\boldsymbol{\Sigma}})^\top\Big] \\
\geq\ &\tilde{\sigma}^2(\mathbf{w}^*)\cdot\Big[\frac{N-1}{N}\widehat{\boldsymbol{\Sigma}} - (\mathbf{S}\mathbf{x}_{i_t}\mathbf{x}_{i_t}^\top\mathbf{S}^\top - \widehat{\boldsymbol{\Sigma}})\Big[\frac{\mathbf{I} - (\mathbf{I}-2\gamma_0\widehat{\boldsymbol{\Sigma}})^{2L_{\mathtt{eff}}}}{N\widehat{\boldsymbol{\Sigma}}}\Big](\mathbf{S}\mathbf{x}_{i_t}\mathbf{x}_{i_t}^\top\mathbf{S}^\top - \widehat{\boldsymbol{\Sigma}})^\top\Big].
\end{aligned}$$

Since $\sup_{x\in[0,1/(2\tilde{\gamma})]}(1-(1-2\tilde{\gamma}x)^{2L_{\mathtt{eff}}})/x \leqslant 4\tilde{\gamma}L_{\mathtt{eff}}$ and $\mathbb{E}_{i_t}[(\mathbf{S}\mathbf{x}_{i_t}\mathbf{x}_{i_t}^\top\mathbf{S}^\top - \widehat{\boldsymbol{\Sigma}})^2] \preceq \mathbb{E}_{i_t}[(\mathbf{S}\mathbf{x}_{i_t}\mathbf{x}_{i_t}^\top\mathbf{S}^\top)^2] \preceq B_{\boldsymbol{\nu}}^2\widehat{\boldsymbol{\Sigma}}$, we further have

$$\begin{aligned}
\mathbb{E}_{\tilde{\boldsymbol{\epsilon}}}\mathbb{E}_{i_t,t\in[L]}[(\boldsymbol{\xi}_{1,i} + \boldsymbol{\xi}_{2,i})(\boldsymbol{\xi}_{1,i} + \boldsymbol{\xi}_{2,i})^\top] &\geq \tilde{\sigma}^2(\mathbf{w}^*)\cdot\Big[\frac{N-1}{N}\widehat{\boldsymbol{\Sigma}} - \frac{4\gamma_0 L_{\mathtt{eff}}}{N}\mathbb{E}_{i_t}[(\mathbf{S}\mathbf{x}_{i_t}\mathbf{x}_{i_t}^\top\mathbf{S}^\top - \widehat{\boldsymbol{\Sigma}})^2]\Big] \\
&\geq \tilde{\sigma}^2(\mathbf{w}^*)\cdot\Big(\frac{N-1}{N} - \frac{4\gamma_0 L_{\mathtt{eff}}}{N}B_{\boldsymbol{\nu}}^2\Big)\widehat{\boldsymbol{\Sigma}} \\
&\geq \tilde{\sigma}^2(\mathbf{w}^*)\cdot\widehat{\boldsymbol{\Sigma}}/2
\end{aligned}$$

when $(L_{\mathtt{eff}}\gamma)B_{\boldsymbol{\nu}}^2/N \leqslant 1/3$.

□

D.3 Fluctuation error under the source condition

Lemma D.5 (Fluctuation error under the source condition)**.** *Under the notation and assumptions in Theorem 3.1 and suppose that $L_{\mathtt{eff}} \lesssim N^{(1-\varepsilon)a}/\gamma$ for some small constant $\varepsilon \in (0,1]$. For any $s \in [0, 1-1/a)$, there exists some (s,ε,a)-dependent constant $c > 0$ such that the* (multi-pass SGD) *process satisfies*

$$\mathbb{E}[\mathsf{Fluc}] \leqslant c\gamma\log N\cdot\Big[1 + \frac{\log^2 N(L_{\mathtt{eff}}\gamma)^{2-s}}{N^2}\Big](L_{\mathtt{eff}}\gamma)^{1/a-1}.$$

with probability at least $1-\exp(-\Omega(M))$ over the randomness of $\mathbf{S}$. Consequently, choosing $s = 1 - 1/(a(1-\varepsilon/2))$ yields

$$\begin{aligned}
\mathbb{E}[\mathsf{Fluc}] &\lesssim \gamma\log N\cdot\Big[1 + \frac{\log^2 N(L_{\mathtt{eff}}\gamma)^{1/(a(1-\varepsilon/2))+1}}{N^2}\Big](L_{\mathtt{eff}}\gamma)^{1/a-1} \\
&\leqslant c'\cdot\gamma\log N\cdot\Big[(L_{\mathtt{eff}}\gamma)^{1/a-1} + \frac{(L_{\mathtt{eff}}\gamma)^{1/a}}{N}\Big]
\end{aligned}$$

for some (ε,a)-dependent constant $c' > 0$ with probability at least $1-\exp(-\Omega(M))$.

Moreover, assume in addition that $L_{\mathtt{eff}} \lesssim N/\gamma$. Then with probability at least $1-\exp(-\Omega(M))$ over the randomness of $\mathbf{S}$, we have

$$\mathbb{E}[\mathsf{Fluc}] \geqslant c''\gamma(L_{\mathtt{eff}}\gamma)^{1/a-1}$$

for some a-dependent constant $c'' > 0$.

Proof of Lemma D.5. The proof follows from instantiating Lemma D.1 and D.4 under the source condition. We start by establishing concentration bounds on some quantities that appear in the bounds in Lemma D.1 and D.4.

First, note that we have for any $s \in [0, 1-1/a)$, conditioned on $\mathbf{S}$ and $\mathbf{w}^*$, with probability at least $1-\delta$ over $(\mathbf{x}_i, y_i)_{i=1}^N$,

$$\max_{i\in[N]}(\mathbf{x}_i^\top\mathbf{S}^\top\mathbf{v}^*)^2 \lesssim \|\mathbf{w}^*\|_{\mathbf{H}}^2\log(N/\delta), \tag{24a}$$

$$\max_{i\in[N]} \tilde{\epsilon}_i^2 \lesssim (\sigma^2 + \|\mathbf{w}^*\|_{\mathbf{H}}^2)\log(N/\delta), \tag{24b}$$

$$\max_{i\in[N]} \|\mathbf{x}_i^\top \mathbf{S}^\top\|_{\mathbf{\Sigma}^{-s}}^2 \lesssim \mathrm{tr}(\mathbf{\Sigma}^{1-s}) + \log(N/\delta) + \sqrt{\mathrm{tr}(\mathbf{\Sigma}^{2-2s})\log(N/\delta)} \lesssim \log(N/\delta), \tag{24c}$$

where Eq. (24a) and (24b) follow from a union bound on concentration inequalities for Gaussian random variables; Eq. (24c) uses Hanson-Wright inequality and Lemma E.5.

Moreover, we will show that conditioned on $\mathbf{S}, \mathbf{w}^*$ and $\boldsymbol{\theta}_t^{(-i)}$, $\mathbf{x}_i^\top \mathbf{S}^\top \boldsymbol{\theta}_t^{(-i)}$ is a zero-mean random Gaussian variable with covariance

$$\mathbb{E}[(\mathbf{x}_i^\top \mathbf{S}^\top \boldsymbol{\theta}_t^{(-i)})^2 \mid \mathbf{S}, \mathbf{w}^*, \boldsymbol{\theta}_t^{(-i)}] = \boldsymbol{\theta}_t^{(-i)\top} \mathbf{\Sigma} \boldsymbol{\theta}_t^{(-i)} \lesssim (T_{\mathbf{B}} + T_{\mathbf{V}}) \cdot (\sigma^2 + \|\mathbf{w}^*\|_{\mathbf{H}}^2), \tag{25}$$

where

$$T_{\mathbf{B}} := \begin{cases} 1 + \log(1/\delta)/N + \mathsf{t}(\delta) \cdot (L_{\mathtt{eff}}\gamma)^2 \cdot (1 + \log(1/\delta)/N)^2 & \text{when} \quad M \leqslant N/2, \\ \mathbf{B}_{\mathbf{B}} \cdot (1 + \log(1/\delta)/N)^2 + 1 & \text{when} \quad M > N/2, \end{cases}$$

$$T_{\mathbf{V}} := \max_{i\in[N]} \|(\widehat{\mathbf{\Sigma}}^{(-i)} + \lambda \mathbf{I})^{-1/2} (\mathbf{\Sigma} + \lambda \mathbf{I})^{1/2}\|^2 \cdot \max_{i\in[N]} \tilde{\epsilon}_i^2 \cdot \widetilde{D}^{\mathsf{U}}/N$$

with $\mathbf{B}_{\mathbf{B}}$ defined in Lemma B.1, $\widetilde{D}^{\mathsf{U}}$ defined in Eq. (17), and $\mathsf{t}(\delta) := 1_{\{\log(1/\delta) \gtrsim N\}}$. Thus, we have by a union bound that

$$\begin{aligned} \max_{i\in[N], t\in[L]} (\mathbf{x}_i^\top \mathbf{S}^\top \boldsymbol{\theta}_t^{(-i)})^2 &\lesssim \max_{i\in[N], t\in[L]} \boldsymbol{\theta}_t^{(-i)\top} \mathbf{\Sigma} \boldsymbol{\theta}_t^{(-i)} \cdot \log(NL/\delta) \\ &\lesssim (T_{\mathbf{B}} + T_{\mathbf{V}}) \cdot (\sigma^2 + \|\mathbf{w}^*\|_{\mathbf{H}}^2) \cdot \log(N/\delta) \end{aligned} \tag{26}$$

with probability at least $1-\delta$ over the randomness of $(\mathbf{x}_i, y_i)_{i=1}^N$ conditioned on $\mathbf{S}$ and $\mathbf{w}^*$. Moreover, we note that $\mu_j(\mathbf{\Sigma}) \eqsim j^{-a}$ for $j \in [M]$ with probability at least $1 - \exp(-\Omega(M))$ by Lemma E.5, and conditioned on $\mathbf{S}$ and $\mathbf{w}^*$, we have $\mu_j(\widehat{\mathbf{\Sigma}}) \eqsim j^{-a}$ for $j \leqslant \min\{M, N/c\}$ and $\mu_j(\widehat{\mathbf{\Sigma}}) \lesssim j^{-a}$ otherwise with probability at least $1 - \exp(-\Omega(N))$ by Lemma E.8.

Proof of the upper bound. Therefore, substituting Eq. (24a)—(24c) and (26) into the expression in $a_{\max}, \mathbf{B}_{\mathbf{\Delta}}$ and F, applying Eq. (16) to bound $\widetilde{D}^{\mathsf{U}}$ (and $\mathrm{tr}(\widehat{\mathbf{\Sigma}}^{1/\alpha})$ for some $\alpha = a + \varepsilon > a$), using part 2 of Lemma D.1[4] and taking expectation w.r.t. $(\mathbf{x}_i, y_i)_{i=1}^N$ conditioned on $\mathbf{S}$ and $\mathbf{w}^*$, it can be verified that

$$\mathbb{E}[\mathsf{Fluc} \mid \mathbf{S}, \mathbf{w}^*] \lesssim (\sigma^2 + \|\mathbf{w}^*\|_{\mathbf{H}}^2) \cdot \gamma \log N \cdot \Big[1 + \frac{\log^2 N (L_{\mathtt{eff}}\gamma)^{2-s}}{N^2}\Big] \cdot (L_{\mathtt{eff}}\gamma)^{1/a-1}.$$

Taking expectation w.r.t. $\mathbf{w}^*$ yields the desired result.

Proof of the lower bound. Setting $s = 0$ in Eq. (24c), we have $B_{\boldsymbol{\xi}} \geqslant 1/2$ when $\gamma L_{\mathtt{eff}} B_{\boldsymbol{\xi}}/N \geqslant 1/3$, which happens with probability at least $1 - N^{-c_1/c_2}$ for some constant $c_1 > 0$ when $\gamma \leqslant c_2/\log N$ for some $c_2 > 0$. Moreover, by the concentration properties on $\mu_j(\widehat{\mathbf{\Sigma}})$ and $\mu_j(\mathbf{\Sigma})$ in the previous discussion, the assumptions on γ, and a union bound, conditioned on $\mathbf{S}$ such that $\mu_j(\mathbf{\Sigma}) \eqsim j^{-a}$ for $j \in [M]$ (which happens with probability at least $1 - e^{-\Omega(M)}$), we have with probability at least $1/2$ over the randomness of $(\mathbf{x}_i, y_i)_{i=1}^N$ that

$$\begin{gathered} \mu_j(\widehat{\mathbf{\Sigma}}) \eqsim j^{-a} \quad \text{for} \quad j \leqslant \min\{M, N/c\}, \ \text{and} \\ \tilde{t} \lesssim (L_{\mathtt{eff}}\gamma)^{1/a}, \quad B_{\boldsymbol{\xi}} \geqslant 1/2, \quad \gamma_0 = \gamma. \end{gathered}$$

Thus, when $(L_{\mathtt{eff}}\gamma)^{1/a} \leqslant M/\tilde{c} \leqslant \min\{M, N/c\}$ for some $\tilde{c}, c > 0$ sufficiently large, we have by Lemma D.4 that

$$\mathbb{E}[\mathsf{Fluc}] \gtrsim \mathbb{E}_{(\mathbf{x}_i)_{i\in[N]}}\Big[\frac{B_{\boldsymbol{\xi}}\gamma_0 L_{\mathtt{eff}}\gamma_0}{10} \cdot \sum_{i>\tilde{t}} \mu_i(\widehat{\mathbf{\Sigma}}) \cdot \mu_i(\mathbf{\Sigma})\Big]$$

[4] More specifically, we apply part 2 of Lemma D.1 on the event with probability at least $1 - \exp(-\Omega(N))$ where conditions on $\widehat{\mathbf{\Sigma}}$ specified in Lemma E.8 hold, and apply part 1 of Lemma D.1 for some $\alpha = a + \varepsilon > 1$ otherwise.

$$\gtrsim L_{\mathtt{eff}}\gamma^2 \cdot \sum_{i=\tilde{t}+1}^{\min\{M,N/c\}} i^{-2a}$$
$$\gtrsim L_{\mathtt{eff}}\gamma^2 \cdot \tilde{t}^{1-2a} \gtrsim \gamma(L_{\mathtt{eff}}\gamma)^{1/a-1}$$

with probability at least $1-e^{-\Omega(M)}$ over the randomness of $\mathbf{S}$.

Proof of claim (25)**.** Note that

$$\boldsymbol{\theta}_t^{(-i)\top}\boldsymbol{\Sigma}\boldsymbol{\theta}_t^{(-i)} \lesssim \mathbf{v}^{*\top}\boldsymbol{\Sigma}\mathbf{v}^* + (\boldsymbol{\theta}_t^{(-i)}-\mathbf{v}^*)^\top\boldsymbol{\Sigma}(\boldsymbol{\theta}_t^{(-i)}-\mathbf{v}^*)$$
$$\overset{(i)}{\lesssim} \|\mathbf{v}^*\|_{\boldsymbol{\Sigma}}^2 + \|\prod_{i=1}^t(\mathbf{I}-\gamma_i\widehat{\boldsymbol{\Sigma}}^{(-i)})\mathbf{v}^*\|_{\boldsymbol{\Sigma}}^2 + \|\mathbf{V}(\widehat{\boldsymbol{\Sigma}}^{(-i)})(\mathbf{S}\mathbf{X}^\top\tilde{\boldsymbol{\epsilon}})^{(-i)}\|_{\boldsymbol{\Sigma}}^2.$$

where step (i) follows from the decomposition

$$\boldsymbol{\theta}_t^{(-i)}-\mathbf{v}^* = \prod_{t=1}^t\Big(\mathbf{I}-\gamma_t\widehat{\boldsymbol{\Sigma}}^{(-i)}\Big)(\boldsymbol{\theta}_0-\mathbf{v}^*) + \mathbf{V}(\widehat{\boldsymbol{\Sigma}}^{(-i)})(\mathbf{S}\mathbf{X}^\top\tilde{\boldsymbol{\epsilon}})^{(-i)}$$
$$= -\prod_{t=1}^t\Big(\mathbf{I}-\gamma_t\widehat{\boldsymbol{\Sigma}}^{(-i)}\Big)\mathbf{v}^* + \mathbf{V}(\widehat{\boldsymbol{\Sigma}}^{(-i)})(\mathbf{S}\mathbf{X}^\top\tilde{\boldsymbol{\epsilon}})^{(-i)}$$

as similar to Eq. (4), where $(\mathbf{S}\mathbf{X}^\top\tilde{\boldsymbol{\epsilon}})^{(-i)} := \sum_{j\neq i}\mathbf{S}\mathbf{x}_j\tilde{\epsilon}_j/N$ and

$$\mathbf{V}(\widehat{\boldsymbol{\Sigma}}^{(-i)}) := \frac{1}{N}\sum_{i=1}^t\gamma_i\cdot\prod_{j=i+1}^t(\mathbf{I}-\gamma_j\widehat{\boldsymbol{\Sigma}}^{(-i)}) = \frac{\mathbf{I}-\prod_{i=1}^t(\mathbf{I}-\gamma_i\widehat{\boldsymbol{\Sigma}}^{(-i)})}{N\widehat{\boldsymbol{\Sigma}}^{(-i)}}.$$

Let $\bar{\mathbf{w}}^* := \boldsymbol{\Sigma}^{1/2}\mathbf{v}^*$. Note that $(\boldsymbol{\theta}_t^{(-i)})_{t=1}^L$ can be viewed as a (GD) process on $(\mathbf{x}_j,y_j)_{j\neq i}$ with stepsize $(N-1)\gamma_t/N$.

Following the proof of Lemma B.1 (Eq. 11 and 12), it can be verified that, conditioned on $\mathbf{S}$ and $\mathbf{w}^*$,

$$\|\prod_{i=1}^t(\mathbf{I}-\gamma_i\widehat{\boldsymbol{\Sigma}}^{(-i)})\mathbf{v}^*\|_{\boldsymbol{\Sigma}}^2 \lesssim \begin{cases} B_{\mathtt{F},1} \overset{d}{=} \|\mathbf{Z}\bar{\mathbf{w}}^*\|_2^2 + (L_{\mathtt{eff}}\gamma)^2\|\mathbf{Z}^\top\mathbf{Z}\|_2^2\mathbf{1}_{\{\mathbf{Z}^\top\mathbf{Z}\not\succeq\mathbf{I}_{M/5}\}}\cdot\|\bar{\mathbf{w}}^*\|_2^2 & \text{when} \quad M\leqslant N/2, \\ B_{\mathtt{F},2} \overset{d}{=} B_{\mathtt{F},3}\cdot\|\mathbf{Z}\bar{\mathbf{w}}^*\|_2^2 + \|\bar{\mathbf{w}}^*\|_2^2 & \text{when} \quad M>N/2, \end{cases}$$

where $\mathbf{Z}\in\mathbb{R}^{(N-1)\times M}$ has i.i.d $\mathcal{N}(0,1/N)$ entries and

$$B_{\mathtt{F},3} := (L_{\mathtt{eff}}\gamma)^2\|\mathbf{Z}_{\tilde{k}:\infty}\boldsymbol{\Sigma}_{\tilde{k}:\infty}\mathbf{Z}_{\tilde{k}:\infty}^\top\|_2^2 + 1 + \|\mathbf{Z}_{\tilde{k}:\infty}\boldsymbol{\Sigma}_{\tilde{k}:\infty}^2\mathbf{Z}_{\tilde{k}:\infty}^\top\|_2$$

with $\tilde{k}=N/2$. In addition, we have $\|\bar{\mathbf{w}}^*\|_2^2 \leqslant \|\mathbf{v}^*\|_{\boldsymbol{\Sigma}}^2 \leqslant \|\mathbf{w}^*\|_{\mathbf{H}}^2$ and $\|\mathbf{Z}\bar{\mathbf{w}}^*\|_2^2 \leqslant \|\mathbf{v}^*\|_{\boldsymbol{\Sigma}}^2\cdot(2+\log(1/\delta)/N)$ with probability at least $1-\delta$ by concentration properties of chi-squared random variables. Therefore, putting pieces together, applying Lemma E.5, Eq. (13) and concentration properties of Gaussian covariance matrices (see e.g., Theorem 6.1 in Wainwright (2019)), we obtain with probability at least $1-\delta$ conditioned on $\mathbf{S}$ and $\mathbf{w}^*$,

$$\|\prod_{i=1}^t(\mathbf{I}-\gamma_i\widehat{\boldsymbol{\Sigma}}^{(-i)})\mathbf{v}^*\|_{\boldsymbol{\Sigma}}^2$$
$$\lesssim \begin{cases} \|\mathbf{w}^*\|_{\mathbf{H}}^2\cdot(2+\log(1/\delta)/N + \mathtt{t}(\delta)\cdot(L_{\mathtt{eff}}\gamma)^2\cdot(2+\log(1/\delta)/N)^2) & \text{when} \quad M\leqslant N/2, \\ \|\mathbf{w}^*\|_{\mathbf{H}}^2\cdot(\mathbf{B}_{\mathbf{B}}\cdot(1+\log(1/\delta)/N)^2+1) & \text{when} \quad M>N/2, \end{cases} \tag{27}$$

where $\mathtt{t}(\delta) := 1_{\{\log(1/\delta)\gtrsim N\}}$ and $\mathbf{B}_{\mathbf{B}}$ is defined in Lemma B.1, with probability at least $1-e^{-\Omega(M)}$ over the randomness of $\mathbf{S}$.

Adopt the shorthand $\mathsf{V}_t^{(-i)}$ for $\mathbf{I}-\prod_{t=1}^L(\mathbf{I}-\gamma_t\widehat{\boldsymbol{\Sigma}})$ and R_i for $\|(\widehat{\boldsymbol{\Sigma}}^{(-i)}+\lambda\mathbf{I})^{-1/2}(\boldsymbol{\Sigma}^{(-i)}+\lambda\mathbf{I})^{1/2}\|^2$. Similarly, following the proof of the upper bound in Lemma C.1, we have (choosing $\lambda=1/(L_{\mathtt{eff}}\gamma)$)

$$\|\mathbf{V}(\widehat{\boldsymbol{\Sigma}}^{(-i)})(\mathbf{S}\mathbf{X}^\top\tilde{\boldsymbol{\epsilon}})^{(-i)}\|_{\boldsymbol{\Sigma}}^2$$

$$
\begin{aligned}
&\lesssim \frac{\max_{i\in[N]}\tilde{\epsilon}_i^2\cdot}{N}\cdot \mathrm{tr}(\mathsf{V}_t^{(-i)}\boldsymbol{\Sigma}\mathsf{V}_t^{(-i)}\widehat{\boldsymbol{\Sigma}}^{-1}) \\
&\leqslant R_i\cdot\frac{\max_{i\in[N]}\tilde{\epsilon}_i^2\cdot}{N}\cdot \mathrm{tr}(\mathsf{V}_t^{(-i),2}+\lambda\widehat{\boldsymbol{\Sigma}}^{-1}\mathsf{V}_t^{(-i),2}) \\
&\overset{(ii)}{\lesssim} \frac{R_i}{N}\cdot\max_{i\in[N]}\tilde{\epsilon}_i^2\cdot\widetilde{D}_i^{\mathsf{U}} \overset{(iii)}{\lesssim} \frac{R_i}{N}\cdot\max_{i\in[N]}\tilde{\epsilon}_i^2\cdot\widetilde{D}^{\mathsf{U}}
\end{aligned}
\tag{28}
$$

where $\widetilde{D}^{\mathsf{U}}$ is defined in Eq. (17) and

$$
\widetilde{D}_i^{\mathsf{U}} := \#\{j\in[M]:\widehat{\lambda}_j^{(-i)}L_{\mathtt{eff}}\gamma_0>1/4\}+(L_{\mathtt{eff}}\gamma_0)\sum_{j:\widehat{\lambda}_j^{(-i)}L_{\mathtt{eff}}\gamma_0\leqslant 1/4}\widehat{\lambda}_j^{(-i)},
$$

and $\widehat{\lambda}_j^{(-i)}$ is the j-th largest eigenvalue of $\widehat{\boldsymbol{\Sigma}}^{(-i)}$. Here, step (ii) uses Eq. (17) and step (iii) uses the fact that $\widehat{\lambda}_j^{(-i)}\leqslant\widehat{\lambda}_j$ for all $j\in[M]$ since $\widehat{\boldsymbol{\Sigma}}^{(-i)}\preceq\widehat{\boldsymbol{\Sigma}}$.

□

D.4 Proof of Lemma D.2

The proof of Lemma D.2 follows from similar ideas as in the proof of Proposition 1 of Pillaud-Vivien et al. (2018). We first state a few lemmas that contribute to the proof. These lemmas are modified versions of the lemmas in Pillaud-Vivien et al. (2018), but we provide their proofs here for completeness.

Lemma D.6 (Semi-stochastic SGD; Lemma 1 in Pillaud-Vivien et al. (2018))**.** *Under the notation and assumptions in Lemma D.2, consider any stochastic process $\widetilde{\boldsymbol{\mu}}_t=(\mathbf{I}-\gamma_t\boldsymbol{\Sigma}_{\boldsymbol{\nu}})\widetilde{\boldsymbol{\mu}}_{t-1}+\gamma_t\cdot\widetilde{\boldsymbol{\xi}}_t$ with $\widetilde{\boldsymbol{\mu}}_0=\mathbf{0}, t\in[L]$ and $(\widetilde{\boldsymbol{\xi}}_t)_{t=1}^L$ such that $\mathbb{E}[\widetilde{\boldsymbol{\xi}}_t]=\mathbf{0}$ and $\mathbb{E}[\widetilde{\boldsymbol{\xi}}_t\widetilde{\boldsymbol{\xi}}_t^\top]\preceq\widetilde{\sigma}_\xi^2\boldsymbol{\Sigma}_{\boldsymbol{\nu}}$. Then for any $u\in[0,1]$, we have*

$$
\mathbb{E}[\|\boldsymbol{\Sigma}_{\boldsymbol{\nu}}^{u/2}\widetilde{\boldsymbol{\mu}}_L\|_2^2]\leqslant c\cdot\widetilde{\sigma}_\xi^2\gamma_0\,\mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\nu}}^{1/\alpha})\cdot(L_{\mathtt{eff}}\gamma_0)^{1/\alpha-u}
$$

for any $\alpha>1$ and some α-dependent constant $c>0$. Moreover, there exists some a-dependent constant $c',\tilde{c}>1$ such that when $\mu_j(\boldsymbol{\Sigma}_{\boldsymbol{\nu}})\eqsim j^{-a}$ for $j\leqslant\min\{M,N/\tilde{c}\}$, we have

$$
\mathbb{E}[\|\boldsymbol{\Sigma}_{\boldsymbol{\nu}}^{u/2}\widetilde{\boldsymbol{\mu}}_L\|_2^2]\leqslant c'\widetilde{\sigma}_\xi^2\cdot\gamma_0(L_{\mathtt{eff}}\gamma_0)^{1/a-u}
$$

for any $u\in[0,1]$.

See the proof of Lemma D.6 in Section D.4.1.

Following the ideas in (Pillaud-Vivien et al., 2018; Aguech et al., 2000), we introduce a sequence of stochastic processes $(\widetilde{\boldsymbol{\mu}}_t^k)_{t=0}^L$ that connects the SGD process in (19) to the semi-stochastic SGD in Lemma D.6. Namely, for $k\geqslant 0$, we define

$$
\widetilde{\boldsymbol{\mu}}_t^k=(\mathbf{I}-\gamma_t\boldsymbol{\Sigma}_{\boldsymbol{\nu}})\widetilde{\boldsymbol{\mu}}_{t-1}^k+\gamma_t\cdot\boldsymbol{\xi}_t^k,\quad\widetilde{\boldsymbol{\mu}}_0^k=\mathbf{0},\quad t\in[L],
\tag{29}
$$

where $\boldsymbol{\xi}_t^0:=\boldsymbol{\xi}_t$ and $\boldsymbol{\xi}_t^k:=(\boldsymbol{\Sigma}_{\boldsymbol{\nu}}-\boldsymbol{\nu}_t\boldsymbol{\nu}_t^\top)\widetilde{\boldsymbol{\mu}}_{t-1}^{k-1}$ for $k\geqslant 1$. It can be verified that

$$
\boldsymbol{\mu}_t-\sum_{i=0}^k\widetilde{\boldsymbol{\mu}}_t^i=(\mathbf{I}-\gamma_t\boldsymbol{\nu}_t\boldsymbol{\nu}_t^\top)\Big(\boldsymbol{\mu}_{t-1}-\sum_{i=0}^{k-1}\widetilde{\boldsymbol{\mu}}_{t-1}^i\Big)+\gamma_t\cdot\boldsymbol{\xi}_t^{k+1}.
$$

Lemma D.7 (Bounds on the covariance; Lemma 2 in Pillaud-Vivien et al. (2018))**.** *Under the notation and assumptions in Lemma D.2 and its proof, for any $k\geqslant 0$, we have*

$$
\mathbb{E}[\boldsymbol{\xi}_t^k\boldsymbol{\xi}_t^{k\top}]\preceq\sigma_\xi^2\gamma_0^kB_{\boldsymbol{\nu}}^{2k}\cdot\boldsymbol{\Sigma}_{\boldsymbol{\nu}}\quad\textit{and}\quad\mathbb{E}[\widetilde{\boldsymbol{\mu}}_t^k\widetilde{\boldsymbol{\mu}}_t^{k\top}]\preceq\sigma_\xi^2\gamma_0^{k+1}B_{\boldsymbol{\nu}}^{2k}\cdot\mathbf{I}.
$$

See the proof of Lemma D.7 in Section D.4.2.

Lemma D.8 (SGD recursion; Lemma 3 in Pillaud-Vivien et al. (2018))**.** *Under the notation and assumptions in Lemma D.2, consider any stochastic process $\widehat{\boldsymbol{\mu}}_t=(\mathbf{I}-\gamma_t\boldsymbol{\nu}_t\boldsymbol{\nu}_t^\top)\widehat{\boldsymbol{\mu}}_{t-1}+\gamma_t\cdot\widehat{\boldsymbol{\xi}}_t$, with $\widehat{\boldsymbol{\mu}}_0=\mathbf{0}, t\in[L]$ and $(\widehat{\boldsymbol{\xi}}_t)_{t=1}^L$ such that $\mathbb{E}[\widehat{\boldsymbol{\xi}}_t]=\mathbf{0}$ and $\mathbb{E}[\widehat{\boldsymbol{\xi}}_t\widehat{\boldsymbol{\xi}}_t^\top]\preceq\widehat{\sigma}_\xi^2\boldsymbol{\Sigma}_{\boldsymbol{\nu}}$. Then*

$$
\mathbb{E}[\|\boldsymbol{\Sigma}_{\boldsymbol{\nu}}^{u/2}\widehat{\boldsymbol{\mu}}_L\|_2^2]\leqslant 2\widehat{\sigma}_\xi^2\cdot\gamma_0^2B_{\boldsymbol{\nu}}^{2u}\,\mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\nu}})L_{\mathtt{eff}}.
$$

for any $u\in[0,1]$.

See the proof of Lemma D.8 in Section D.4.3.

With these lemmas at hand, we are ready to prove Lemma D.2. Performing a decomposition as in the proof of Proposition 1 in Pillaud-Vivien et al. (2018) and using Lemma D.6 on $\widetilde{\boldsymbol{\mu}}_t^i$ for $i \in [0, k]$ and D.8 on $\boldsymbol{\mu}_L - \sum_{i=0}^k \widetilde{\boldsymbol{\mu}}_L^i$, we find

$$
\begin{aligned}
&(\mathbb{E}[\|\boldsymbol{\Sigma}_{\boldsymbol{\nu}}^{u/2}\boldsymbol{\mu}_L\|_2^2])^{1/2} \\
&\leqslant \sum_{i=0}^{k}(\mathbb{E}[\|\boldsymbol{\Sigma}_{\boldsymbol{\nu}}^{u/2}\widetilde{\boldsymbol{\mu}}_L^i\|_2^2])^{1/2} + (\mathbb{E}[\|\boldsymbol{\Sigma}_{\boldsymbol{\nu}}^{u/2}(\boldsymbol{\mu}_L - \sum_{i=0}^{k}\widetilde{\boldsymbol{\mu}}_L^i)\|_2^2])^{1/2} \\
&\lesssim \sum_{i=0}^{k}(\sigma_\xi^2\gamma_0^i B_{\boldsymbol{\nu}}^{2i}\cdot\gamma_0\operatorname{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\nu}}^{1/\alpha}){L_{\mathtt{eff}}}^{1/\alpha-u})^{1/2} + (\sigma_\xi^2\gamma_0^{k+1}B_{\boldsymbol{\nu}}^{2k+2+2u}\cdot\gamma_0^2\operatorname{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\nu}})L_{\mathtt{eff}})^{1/2} \\
&\leqslant (\sigma_\xi^2\cdot\gamma_0\operatorname{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\nu}}^{1/\alpha})(L_{\mathtt{eff}}\gamma_0)^{1/\alpha-u})^{1/2}\cdot\sum_{i=0}^{k}(\gamma_0 B_{\boldsymbol{\nu}}^2)^{i/2} + (\sigma_\xi^2\gamma_0^{k+3}B_{\boldsymbol{\nu}}^{2k+2+2u}\cdot\operatorname{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\nu}})L_{\mathtt{eff}})^{1/2} \\
&\leqslant 2(\sigma_\xi^2\cdot\gamma_0\operatorname{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\nu}}^{1/\alpha})(L_{\mathtt{eff}}\gamma_0)^{1/\alpha-u})^{1/2} + (\sigma_\xi^2\gamma_0^{k+3}B_{\boldsymbol{\nu}}^{2k+2+2u}\cdot\operatorname{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\nu}})L_{\mathtt{eff}})^{1/2},
\end{aligned}
$$

where the last inequality follows as $\gamma_0 B_{\boldsymbol{\nu}}^2 \leqslant 1/4$ by the assumption in Lemma D.2. Finally, letting $k \to \infty$ and noting that $\sigma_\xi^2\gamma_0^{k+3}B_{\boldsymbol{\nu}}^{2k+2+2u}\cdot\operatorname{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\nu}})L_{\mathtt{eff}} \overset{k\to\infty}{\longrightarrow} 0$, we obtain the desired result. The second part of Lemma D.2 follows from similar arguments and therefore we omit the proof.

D.4.1 Proof of Lemma D.6

By definition of $\widetilde{\boldsymbol{\mu}}_t$, we have

$$
\widetilde{\boldsymbol{\mu}}_L = \sum_{t=1}^{L}\gamma_t\cdot\prod_{i=t+1}^{L}(\mathbf{I}-\gamma_i\boldsymbol{\Sigma}_{\boldsymbol{\nu}})\boldsymbol{\xi}_t.
$$

Thus,

$$
\begin{aligned}
&\mathbb{E}[\|\boldsymbol{\Sigma}_{\boldsymbol{\nu}}^{u/2}\widetilde{\boldsymbol{\mu}}_L\|_2^2] \\
&= \mathbb{E}[\|\boldsymbol{\Sigma}_{\boldsymbol{\nu}}^{u/2}\sum_{t=1}^{L}\gamma_t\cdot\prod_{i=t+1}^{L}(\mathbf{I}-\gamma_i\boldsymbol{\Sigma}_{\boldsymbol{\nu}})\boldsymbol{\xi}_t\|_2^2] = \sum_{t=1}^{L}\gamma_t^2\cdot\operatorname{tr}(\mathbb{E}[\boldsymbol{\Sigma}_{\boldsymbol{\nu}}^{u}\prod_{i=t+1}^{L}(\mathbf{I}-\gamma_i\boldsymbol{\Sigma}_{\boldsymbol{\nu}})^2\boldsymbol{\xi}_t\boldsymbol{\xi}_t^\top]) \\
&\leqslant \widetilde{\sigma}_\xi^2\cdot\sum_{t=1}^{L}\gamma_t^2\cdot\operatorname{tr}(\prod_{i=t+1}^{L}(\mathbf{I}-\gamma_i\boldsymbol{\Sigma}_{\boldsymbol{\nu}})^2\boldsymbol{\Sigma}_{\boldsymbol{\nu}}^{1+u}) \\
&= \widetilde{\sigma}_\xi^2\cdot\sum_{k=0}^{\lfloor\log L\rfloor-1}\gamma_{L_{\mathtt{eff}}k+1}^2\cdot\operatorname{tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{\nu}}^{1+u}\cdot\frac{\mathbf{I}-(\mathbf{I}-\gamma_{L_{\mathtt{eff}}k+1}\boldsymbol{\Sigma}_{\boldsymbol{\nu}})^{2L_{\mathtt{eff}}}}{2\gamma_{L_{\mathtt{eff}}k+1}\boldsymbol{\Sigma}_{\boldsymbol{\nu}}-(\gamma_{L_{\mathtt{eff}}k+1}\boldsymbol{\Sigma}_{\boldsymbol{\nu}})^2}\cdot\prod_{j=k+1}^{\lfloor\log L\rfloor-1}(\mathbf{I}-\gamma_{L_{\mathtt{eff}}j+1}\boldsymbol{\Sigma}_{\boldsymbol{\nu}})^{2L_{\mathtt{eff}}}\right) \\
&\leqslant \widetilde{\sigma}_\xi^2\cdot\sum_{k=0}^{\lfloor\log L\rfloor-2}\gamma_{L_{\mathtt{eff}}k+1}\operatorname{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\nu}}^{u}\cdot(\mathbf{I}-(\mathbf{I}-\gamma_{L_{\mathtt{eff}}k+1}\boldsymbol{\Sigma}_{\boldsymbol{\nu}})^{2L_{\mathtt{eff}}})(\mathbf{I}-\gamma_{L_{\mathtt{eff}}(k+1)+1}\boldsymbol{\Sigma}_{\boldsymbol{\nu}})^{2L_{\mathtt{eff}}}) \\
&\quad + \operatorname{tr}(\widetilde{\sigma}_\xi^2\cdot\frac{\gamma_0}{L}\boldsymbol{\Sigma}_{\boldsymbol{\nu}}^{u}\cdot(\mathbf{I}-(\mathbf{I}-\gamma_0\boldsymbol{\Sigma}_{\boldsymbol{\nu}}/L)^{2L_{\mathtt{eff}}})),
\end{aligned}
\tag{30}
$$

where the first inequality uses $\mathbb{E}[\boldsymbol{\xi}_t\boldsymbol{\xi}_t^\top] \preceq \widetilde{\sigma}_\xi^2\boldsymbol{\Sigma}_{\boldsymbol{\nu}}$.

Part 1 of Lemma D.6. Comtinuing the calculation in Eq. (30), we have

$$
\begin{aligned}
&\mathbb{E}[\|\boldsymbol{\Sigma}_{\boldsymbol{\nu}}^{u/2}\widetilde{\boldsymbol{\mu}}_L\|_2^2] \\
&\leqslant \widetilde{\sigma}_\xi^2\cdot\operatorname{tr}\Big[\sum_{k=0}^{\lfloor\log L\rfloor-2}\frac{\gamma_{L_{\mathtt{eff}}k+1}}{(2\gamma_{L_{\mathtt{eff}}(k+1)+1}L_{\mathtt{eff}})^u}\cdot(\mathbf{I}-(\mathbf{I}-\gamma_{L_{\mathtt{eff}}k+1}\boldsymbol{\Sigma}_{\boldsymbol{\nu}})^{2L_{\mathtt{eff}}}) + \frac{\gamma_0}{L}\boldsymbol{\Sigma}_{\boldsymbol{\nu}}\cdot(\mathbf{I}-(\mathbf{I}-\gamma_0\boldsymbol{\Sigma}_{\boldsymbol{\nu}}/L)^{2L_{\mathtt{eff}}})\Big] \\
&\lesssim \widetilde{\sigma}_\xi^2\cdot\Big[\sum_{k=0}^{\lfloor\log L\rfloor-2}\frac{\gamma_{L_{\mathtt{eff}}k+1}^{1-u}}{{L_{\mathtt{eff}}}^u}\cdot\operatorname{tr}((\gamma_{L_{\mathtt{eff}}k+1}L_{\mathtt{eff}}\boldsymbol{\Sigma}_{\boldsymbol{\nu}})^{1/\alpha}) + \frac{\gamma_0}{L}\cdot\operatorname{tr}((\gamma_0\boldsymbol{\Sigma}_{\boldsymbol{\nu}})^{1/\alpha})\Big]
\end{aligned}
$$

$$\lesssim \widetilde{\sigma}_\xi^2 \cdot \gamma_0^{1-u+1/\alpha} \operatorname{tr}(\mathbf{\Sigma}_{\boldsymbol{\nu}}^{1/\alpha}) {L_{\mathtt{eff}}}^{1/\alpha-u} \lesssim \widetilde{\sigma}_\xi^2 \cdot \gamma_0 \operatorname{tr}(\mathbf{\Sigma}_{\boldsymbol{\nu}}^{1/\alpha})(L_{\mathtt{eff}}\gamma_0)^{1/\alpha-u},$$

where the first inequality uses $\sup_{x\in[0,1/\gamma_0]} x^u(1-\gamma_0 x)^{2L_{\mathtt{eff}}} \leqslant 1/[2\gamma_0 L_{\mathtt{eff}}]^u$ for any $u \in [0,1]$, the second inequality follows from $1-(1-\gamma_0 x)^{2L_{\mathtt{eff}}} \leqslant (2\gamma_0 L_{\mathtt{eff}} x)^{1/\alpha}$ for any $\alpha > 1$ and $x \in [0, 1/\gamma_0]$ by Bernoulli's inequality, and the last inequality follows from the stepsize definition (2). This gives the first part of Lemma D.6.

Part 2 of Lemma D.6. Similarly, continuing the calculation in Eq. (30) and noting that $\sup_{x\in[0,1/\gamma]}[1-(1-\gamma x)^{2L_{\mathtt{eff}}}]/x \leqslant 2\gamma L_{\mathtt{eff}}$, we obtain

$$\mathbb{E}[\|\mathbf{\Sigma}_{\boldsymbol{\nu}}^{u/2}\widetilde{\boldsymbol{\mu}}_L\|_2^2] \leqslant \widetilde{\sigma}_\xi^2 \cdot \operatorname{tr}\Big[\sum_{k=0}^{\lfloor \log L\rfloor-2} 2L_{\mathtt{eff}}\gamma_{L_{\mathtt{eff}}k+1}^2 \mathbf{\Sigma}_{\boldsymbol{\nu}}^{1+u} \cdot (\mathbf{I}-\gamma_{L_{\mathtt{eff}}(k+1)+1}\mathbf{\Sigma}_{\boldsymbol{\nu}})^{2L_{\mathtt{eff}}} + \frac{2\gamma_0^2}{L}\mathbf{\Sigma}_{\boldsymbol{\nu}}^{1+u}\Big]. \tag{31}$$

Denote the eigenvalues of $\mathbf{\Sigma}_{\boldsymbol{\nu}}$ by $\widehat{\lambda}_1 \geqslant \widehat{\lambda}_2 \geqslant \ldots \geqslant \widehat{\lambda}_N$ (let $\widehat{\lambda}_j = 0$ for $j > M$). Choose $\iota = (L_{\mathtt{eff}}\gamma)^{1/a}$. When $M \leqslant N/\tilde{c}$, we have $\mu_j(\mathbf{\Sigma}_{\boldsymbol{\nu}}) \eqsim j^{-a}$ for $j \leqslant M$ and otherwise 0. When $M > N/\tilde{c}$, we have $\iota \leqslant N/\tilde{c}$, $\widehat{\lambda}_j \eqsim j^{-a}$ for $j \leqslant N/\tilde{c}$ and otherwise $\widehat{\lambda}_j \lesssim j^{-a}$ by monotonicity of $\widehat{\lambda}_j$. In both cases, we have $\widehat{\lambda}_j \eqsim j^{-a}$ (or $\widehat{\lambda}_j = 0$) for $j \leqslant \iota$ and $\widehat{\lambda}_j \lesssim j^{-a}$ for $j > \iota$.

Since $f(x) > f(0)$ for any $x \in [0, 1/\tilde{\gamma}]$ and $f(x) = x^{1+u}(1-\tilde{\gamma}x/2)^{2L_{\mathtt{eff}}}$ for any $u \in [0,1]$, to obtain an upper bound on $\operatorname{tr}(\mathbf{\Sigma}_{\boldsymbol{\nu}}^{1+u} \cdot (\mathbf{I}-\tilde{\gamma}\mathbf{\Sigma}_{\boldsymbol{\nu}}/2)^{2L_{\mathtt{eff}}})$, we can w.l.o.g. assume $\widehat{\lambda}_j \eqsim j^{-a}$ for $j \leqslant \iota$ and $\widehat{\lambda}_j \lesssim j^{-a}$ for $j > \iota$. Under this assumption, for any $\tilde{\gamma} \in [0, 1/(4\widehat{\lambda}_1)]$,

$$\begin{aligned}
&\operatorname{tr}(\mathbf{\Sigma}_{\boldsymbol{\nu}}^{1+u} \cdot (\mathbf{I}-\tilde{\gamma}\mathbf{\Sigma}_{\boldsymbol{\nu}}/2)^{2L_{\mathtt{eff}}}) \\
\lesssim& \sum_{j=1}^N \widehat{\lambda}_j^{1+u} \cdot (\mathbf{I}-\tilde{\gamma}\widehat{\lambda}_j/2)^{2L_{\mathtt{eff}}} \\
\lesssim& \sum_{j>\iota} \widehat{\lambda}_j^{1+u} \cdot (\mathbf{I}-\tilde{\gamma}\widehat{\lambda}_j/2)^{2L_{\mathtt{eff}}} + \sum_{k=0}^\infty \sum_{j\in[\iota/2^{k+1},\iota/2^k)} \widehat{\lambda}_j^{1+u} \cdot (\mathbf{I}-\tilde{\gamma}\widehat{\lambda}_j/2)^{2L_{\mathtt{eff}}} \\
\lesssim& \iota^{1-(1+u)a} + \sum_{k=0}^{\lfloor \log \iota\rfloor+1} \sum_{j\in[\iota/2^{k+1},\iota/2^k)} \widehat{\lambda}_j^{1+u} \cdot (1-\frac{2^{ka}}{2L_{\mathtt{eff}}})^{2L_{\mathtt{eff}}} \\
\lesssim& \iota^{1-(1+u)a} + \sum_{k=0}^\infty (\iota/2^k)^{1-(1+u)a} \cdot e^{-2^{ka}} \lesssim \iota^{1-(1+u)a} \cdot (1+\sum_{k=0}^\infty 2^{k((1+u)a-1)} e^{-2^{ka}}) \\
\lesssim& \iota^{1-(1+u)a} = (L_{\mathtt{eff}}\tilde{\gamma})^{1-(1+u)a},
\end{aligned} \tag{32}$$

and $\gamma_0^2 \operatorname{tr}(\mathbf{\Sigma}_{\boldsymbol{\nu}}^{1+u})/L \lesssim \gamma^2/L_{\mathtt{eff}}$. Substituting these into Eq. (31) yields

$$\begin{aligned}
\mathbb{E}[\|\mathbf{\Sigma}_{\boldsymbol{\nu}}^{u/2}\widetilde{\boldsymbol{\mu}}_L\|_2^2] &\lesssim \widetilde{\sigma}_\xi^2 \cdot \Big[\sum_{k=0}^{\lfloor \log L\rfloor-2} L_{\mathtt{eff}}\gamma_{L_{\mathtt{eff}}k+1}^2 \cdot (L_{\mathtt{eff}}\gamma_{L_{\mathtt{eff}}k+1})^{1/a-u-1} + \gamma_0^2/L_{\mathtt{eff}}\Big] \\
&\lesssim \widetilde{\sigma}_\xi^2 \cdot \gamma_0 \cdot (L_{\mathtt{eff}}\gamma_0)^{1/a-u}.
\end{aligned}$$

D.4.2 Proof of Lemma D.7

We prove this lemma by induction. When $k = 0$, we have $\mathbb{E}[\boldsymbol{\xi}_t^0 \boldsymbol{\xi}_t^{0\top}] = \mathbb{E}[\boldsymbol{\xi}_t \boldsymbol{\xi}_t^\top] \preceq \sigma_\xi^2 \mathbf{\Sigma}_{\boldsymbol{\nu}}$ and

$$\begin{aligned}
\mathbb{E}[\widetilde{\boldsymbol{\mu}}_t^0 \widetilde{\boldsymbol{\mu}}_t^{0\top}] &= \sum_{i=1}^t \gamma_i^2 \prod_{j=i+1}^t (\mathbf{I}-\gamma_j\mathbf{\Sigma}_{\boldsymbol{\nu}})\mathbb{E}[\boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top] \prod_{j=i+1}^t (\mathbf{I}-\gamma_j\mathbf{\Sigma}_{\boldsymbol{\nu}}) \\
&\preceq \sigma_\xi^2\gamma_0 \sum_{i=1}^t \gamma_t \cdot \prod_{j=i+1}^t (\mathbf{I}-\gamma_j\mathbf{\Sigma}_{\boldsymbol{\nu}})\mathbf{\Sigma}_{\boldsymbol{\nu}} \preceq \sigma_\xi^2\gamma_0\Big(\mathbf{I}-\prod_{i=1}^t(\mathbf{I}-\gamma_i\mathbf{\Sigma}_{\boldsymbol{\nu}})\Big) \lesssim \sigma_\xi^2\gamma_0\mathbf{I}.
\end{aligned}$$

Now, assume the lemma holds for some $k \geqslant 0$, we show that it also holds for $k+1$. For $\boldsymbol{\xi}_t^{k+1}$, we have

$$\mathbb{E}[\boldsymbol{\xi}_t^{k+1}\boldsymbol{\xi}_t^{k+1\top}] \preceq \mathbb{E}[(\mathbf{\Sigma}_{\boldsymbol{\nu}}-\boldsymbol{\nu}_t\boldsymbol{\nu}_t^\top)\mathbb{E}[\widetilde{\boldsymbol{\mu}}_{t-1}^k\widetilde{\boldsymbol{\mu}}_{t-1}^{k\top}](\mathbf{\Sigma}_{\boldsymbol{\nu}}-\boldsymbol{\nu}_t\boldsymbol{\nu}_t^\top)] \preceq \sigma_\xi^2\gamma_0^{k+1}B_{\boldsymbol{\nu}}^{2k} \cdot \mathbb{E}[(\mathbf{\Sigma}_{\boldsymbol{\nu}}-\boldsymbol{\nu}_t\boldsymbol{\nu}_t^\top)^2]$$

$$\leq \sigma_\xi^2 \gamma_0^{k+1} B_{\boldsymbol{\nu}}^{2k} \cdot \mathbb{E}[\boldsymbol{\nu}_t \boldsymbol{\nu}_t^\top \boldsymbol{\nu}_t \boldsymbol{\nu}_t^\top] \leq \sigma_\xi^2 \gamma_0^{k+1} B_{\boldsymbol{\nu}}^{2(k+1)} \cdot \boldsymbol{\Sigma}_{\boldsymbol{\nu}}.$$

For $\widetilde{\boldsymbol{\mu}}_t^{k+1}$, we have

$$\begin{aligned}
\mathbb{E}[\widetilde{\boldsymbol{\mu}}_t^{k+1} \widetilde{\boldsymbol{\mu}}_t^{k+1\top}] &= \sum_{i=1}^t \gamma_i^2 \prod_{j=i+1}^t (\mathbf{I} - \gamma_j \boldsymbol{\Sigma}_{\boldsymbol{\nu}}) \mathbb{E}[\boldsymbol{\xi}_i^{k+1} \boldsymbol{\xi}_i^{k+1\top}] \prod_{j=i+1}^t (\mathbf{I} - \gamma_j \boldsymbol{\Sigma}_{\boldsymbol{\nu}}) \\
&\leq \sigma_\xi^2 \gamma_0^{k+2} B_{\boldsymbol{\nu}}^{2(k+1)} \sum_{i=1}^t \gamma_t \cdot \prod_{j=i+1}^t (\mathbf{I} - \gamma_j \boldsymbol{\Sigma}_{\boldsymbol{\nu}}) \boldsymbol{\Sigma}_{\boldsymbol{\nu}} \leq \sigma_\xi^2 \gamma_0^{k+2} B_{\boldsymbol{\nu}}^{2(k+1)} \cdot \mathbf{I}.
\end{aligned}$$

This completes the induction.

D.4.3 Proof of Lemma D.8

By definition of $\widehat{\boldsymbol{\mu}}_t$, we have

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{\Sigma}_{\boldsymbol{\nu}}^{u/2} \widehat{\boldsymbol{\mu}}_L\|_2^2] &= \mathbb{E}[\|\boldsymbol{\Sigma}_{\boldsymbol{\nu}}^{u/2} \sum_{t=1}^L \gamma_t \cdot \prod_{i=t+1}^L (\mathbf{I} - \gamma_i \boldsymbol{\nu}_i \boldsymbol{\nu}_i^\top) \widehat{\boldsymbol{\xi}}_t \|_2^2] \\
&= \sum_{t=1}^L \gamma_t^2 \cdot \operatorname{tr}\Big(\mathbb{E}\Big[\boldsymbol{\Sigma}_{\boldsymbol{\nu}}^u \prod_{i=t+1}^L (\mathbf{I} - \gamma_i \boldsymbol{\nu}_i \boldsymbol{\nu}_i^\top) \widehat{\boldsymbol{\xi}}_t \widehat{\boldsymbol{\xi}}_t^\top \prod_{i=t+1}^L (\mathbf{I} - \gamma_i \boldsymbol{\nu}_i \boldsymbol{\nu}_i^\top) \Big] \Big) \\
&\leqslant \widehat{\sigma}_\xi^2 \sum_{t=1}^L \gamma_t^2 \cdot \operatorname{tr}\Big(\boldsymbol{\Sigma}_{\boldsymbol{\nu}}^u \prod_{i=t+1}^L (\mathbf{I} - \gamma_i \boldsymbol{\nu}_i \boldsymbol{\nu}_i^\top) \boldsymbol{\Sigma}_{\boldsymbol{\nu}} \prod_{i=t+1}^L (\mathbf{I} - \gamma_i \boldsymbol{\nu}_i \boldsymbol{\nu}_i^\top) \Big) \\
&\leqslant \widehat{\sigma}_\xi^2 \cdot \sum_{t=1}^L \gamma_t^2 \operatorname{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\nu}}) \|\boldsymbol{\Sigma}_{\boldsymbol{\nu}}\|_2^u \leqslant 2\widehat{\sigma}_\xi^2 B_{\boldsymbol{\nu}}^{2u} \cdot \gamma_0^2 \operatorname{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\nu}}) L_{\mathtt{eff}},
\end{aligned}$$

where the last inequality follows since $\boldsymbol{\Sigma}_{\boldsymbol{\nu}}^2 \preceq \mathbb{E}[\boldsymbol{\nu}_t \boldsymbol{\nu}_t^\top \boldsymbol{\nu}_t \boldsymbol{\nu}_t^\top] \preceq B_{\boldsymbol{\nu}}^2 \boldsymbol{\Sigma}_{\boldsymbol{\nu}}$ and $\sum_{t=1}^L \gamma_t^2 \leqslant 2L_{\mathtt{eff}} \gamma_0^2$.

D.5 Proof of Lemma D.3

Let $\boldsymbol{\Delta}_t^{(-i)} := \boldsymbol{\theta}_t - \boldsymbol{\theta}_t^{(-i)}$. For any $i \in [N], t \in [L]$, we have

$$\begin{aligned}
(\mathbf{x}_i^\top \mathbf{S}^\top \boldsymbol{\theta}_t)^2 &\lesssim 2(\mathbf{x}_i^\top \mathbf{S}^\top \boldsymbol{\theta}_t^{(-i)})^2 + 2(\mathbf{x}_i^\top \mathbf{S}^\top \boldsymbol{\Delta}_t^{(-i)})^2 \leqslant 2(\mathbf{x}_i^\top \mathbf{S}^\top \boldsymbol{\theta}_t^{(-i)})^2 + 2\|\mathbf{x}_i^\top \mathbf{S}^\top\|_{\boldsymbol{\Sigma}^{-s}}^2 \cdot \|\boldsymbol{\Delta}_t^{(-i)}\|_{\boldsymbol{\Sigma}^s}^2 \\
&\lesssim \max_{i \in [N], t \in [L]} (\mathbf{x}_i^\top \mathbf{S}^\top \boldsymbol{\theta}_t^{(-i)})^2 + \max_{i \in [N]} \|\mathbf{x}_i^\top \mathbf{S}^\top\|_{\boldsymbol{\Sigma}^{-s}}^2 \cdot \max_{i \in [N], t \in [L]} \|\boldsymbol{\Delta}_t^{(-i)}\|_{\boldsymbol{\Sigma}^s}^2
\end{aligned}$$

It remains to bound $\max_{i \in [N], t \in [L]} \|\boldsymbol{\Delta}_t^{(-i)}\|_{\boldsymbol{\Sigma}^s}^2$. Adopt the shorthand notation $a_t^{(-i)} := y_i + \mathbf{x}_i^\top \mathbf{S}^\top \boldsymbol{\theta}_{t-1}^{(-i)}$ and recell $a_{\max} = \max_{i \in [N], t \in [L]} |a_t^{(-i)}|$. By taking the difference between the (GD) process and the (LOO-GD) process, we have

$$\boldsymbol{\Delta}_t^{(-i)} = (\mathbf{I} - \gamma_t \widehat{\boldsymbol{\Sigma}}) \boldsymbol{\Delta}_{t-1}^{(-i)} + \frac{\gamma_t a_t^{(-i)}}{N} \mathbf{S} \mathbf{x}_i = \underbrace{\Big[\sum_{i=1}^t \frac{\gamma_i a_t^{(-i)}}{N} \cdot \prod_{j=i+1}^t (\mathbf{I} - \gamma_j \widehat{\boldsymbol{\Sigma}}) \Big]}_{:= \mathsf{V}_{i,t}} \mathbf{S} \mathbf{x}_i.$$

Therefore, for $\lambda = 1/(L_{\mathtt{eff}} \gamma)$,

$$\begin{aligned}
\|\boldsymbol{\Delta}_t^{(-i)}\|_{\boldsymbol{\Sigma}^s}^2 &= \operatorname{tr}(\mathbf{x}_i^\top \mathbf{S}^\top \mathsf{V}_{i,t} \boldsymbol{\Sigma}^s \mathsf{V}_{i,t}^\top \mathbf{S} \mathbf{x}_i) \\
&= \operatorname{tr}(\mathbf{x}_i^\top \mathbf{S}^\top \mathsf{V}_{i,t} \boldsymbol{\Sigma}^s [\boldsymbol{\Sigma} + \lambda \mathbf{I}]^{1/2} [\boldsymbol{\Sigma} + \lambda \mathbf{I}]^{-1/2} \boldsymbol{\Sigma}^s [\boldsymbol{\Sigma} + \lambda \mathbf{I}]^{-1/2} [\boldsymbol{\Sigma} + \lambda \mathbf{I}]^{1/2} \mathsf{V}_{i,t} \mathbf{S} \mathbf{x}_i) \\
&\leqslant \sup_{x \geqslant 0} \frac{x^s}{x + \lambda} \cdot \operatorname{tr}(\mathbf{x}_i^\top \mathbf{S}^\top \mathsf{V}_{i,t} (\boldsymbol{\Sigma} + \lambda \mathbf{I}) \mathsf{V}_{i,t} \mathbf{S} \mathbf{x}_i) \\
&\lesssim \lambda^{s-1} \cdot \|(\widehat{\boldsymbol{\Sigma}} + \lambda \mathbf{I})^{-1/2} (\boldsymbol{\Sigma} + \lambda \mathbf{I})^{1/2}\|^2 \cdot \operatorname{tr}(\mathbf{x}_i^\top \mathbf{S}^\top \mathsf{V}_{i,t} (\widehat{\boldsymbol{\Sigma}} + \lambda \mathbf{I}) \mathsf{V}_{i,t} \mathbf{S} \mathbf{x}_i).
\end{aligned}$$

Note that

$$\mathsf{V}_{i,t} (\widehat{\boldsymbol{\Sigma}} + \lambda \mathbf{I}) \mathsf{V}_{i,t} \preceq \mathsf{V}_{\max} (\widehat{\boldsymbol{\Sigma}} + \lambda \mathbf{I}) \mathsf{V}_{\max},$$

where

$$\mathsf{V}_{\max} := \sum_{i=1}^{t} \frac{\gamma_i a_{\max}}{N} \cdot \prod_{j=i+1}^{t} (\mathbf{I} - \gamma_j \widehat{\mathbf{\Sigma}}) = a_{\max} \cdot \frac{\mathbf{I} - \prod_{i=1}^{t}(\mathbf{I} - \gamma_i \widehat{\mathbf{\Sigma}})}{N\widehat{\mathbf{\Sigma}}}.$$

Adopt the shorthand notation $\mathsf{V}_t = \mathbf{I} - \prod_{i=1}^{t}(\mathbf{I} - \gamma_i \widehat{\mathbf{\Sigma}})$. Choosing $\lambda = 1/(L_{\mathtt{eff}}\gamma)$ in the last display and taking the supremum over $t \in [L], i \in [N]$, we obtain

$$\begin{aligned}
\max_{i\in[N],t\in[L]} \|\mathbf{\Delta}_t^{(-i)}\|_{\mathbf{\Sigma}^s}^2 &\lesssim \frac{\|(\widehat{\mathbf{\Sigma}} + \lambda\mathbf{I})^{-1/2}(\mathbf{\Sigma} + \lambda\mathbf{I})^{1/2}\|^2}{(L_{\mathtt{eff}}\gamma)^{s-1}} \cdot \frac{a_{\max}^2}{N^2} \cdot \mathrm{tr}(\mathbf{x}_i^\top \mathbf{S}^\top \widehat{\mathbf{\Sigma}}^{-1} \mathsf{V}_t (\widehat{\mathbf{\Sigma}} + \lambda\mathbf{I}) \mathsf{V}_t \widehat{\mathbf{\Sigma}}^{-1} \mathbf{S}\mathbf{x}_i) \\
&\lesssim \frac{\|(\widehat{\mathbf{\Sigma}} + \lambda\mathbf{I})^{-1/2}(\mathbf{\Sigma} + \lambda\mathbf{I})^{1/2}\|^2}{(L_{\mathtt{eff}}\gamma)^{s-1}} \cdot \frac{a_{\max}^2}{N^2} \cdot \max_{i\in[N]} \|\mathbf{S}\mathbf{x}_i\|_2^2 \cdot \|\widehat{\mathbf{\Sigma}}^{-1} \mathsf{V}_t (\widehat{\mathbf{\Sigma}} + \lambda\mathbf{I}) \mathsf{V}_t \widehat{\mathbf{\Sigma}}^{-1}\|_2.
\end{aligned}$$

Moreover, we have

$$\|\widehat{\mathbf{\Sigma}}^{-1} \mathsf{V}_t (\widehat{\mathbf{\Sigma}} + \lambda\mathbf{I}) \mathsf{V}_t \widehat{\mathbf{\Sigma}}^{-1}\| \leqslant \|\mathsf{V}_t^2 \widehat{\mathbf{\Sigma}}^{-1}\| + \lambda \cdot \|\widehat{\mathbf{\Sigma}}^{-1} \mathsf{V}_t\|^2 \overset{(i)}{\leqslant} L_{\mathtt{eff}}\gamma,$$

where step (i) follows from $\|\mathsf{V}_t\| \leqslant 1$ and $\sup_{x\in[0,1/\gamma]}(1 - \prod_{i=1}^{t}(1 - \gamma_i x)) \lesssim L_{\mathtt{eff}}\gamma$ by the stepsize definition (2). Combining the last two displays, we find

$$\max_{i\in[N],t\in[L]} \|\mathbf{\Delta}_t^{(-i)}\|_{\mathbf{\Sigma}^s}^2 \lesssim a_{\max}^2 \cdot \max_{i\in[N]} \|\mathbf{S}\mathbf{x}_i\|_2^2 \cdot \|(\widehat{\mathbf{\Sigma}} + \lambda\mathbf{I})^{-1/2}(\mathbf{\Sigma} + \lambda\mathbf{I})^{1/2}\|^2 \cdot \frac{(L_{\mathtt{eff}}\gamma)^{2-s}}{N^2}.$$

This completes the proof.

E Auxiliary lemmas

In this section, we provide some auxiliary lemmas that are used in the proofs.

E.1 General concentration bounds

Lemma E.1. *Let $\boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \ldots, \boldsymbol{\nu}_N$ be i.i.d. samples from $\mathcal{N}(0, \Sigma)$ for some $\Sigma \in \mathbb{R}^{p\times p}$. Let $\widehat{\Sigma} = \sum_{i=1}^N \boldsymbol{\nu}_i\boldsymbol{\nu}_i^\top/N$. Assume that $\sum_{i=1}^p \frac{\mu_i(\Sigma)}{\mu_i(\Sigma)+\lambda} \leqslant N/4$. Then with probability at least $1 - e^{-\Omega(N)}$*

$$\|(\widehat{\Sigma} + \lambda\mathbf{I}_p)^{-1/2}\Sigma^{1/2}\|_2 \leqslant \|(\widehat{\Sigma} + \lambda\mathbf{I}_p)^{-1/2}(\Sigma + \lambda\mathbf{I}_p)^{1/2}\|_2 \leqslant 3.$$

Moreover, the expectation $\mathbb{E}\|(\widehat{\Sigma} + \lambda\mathbf{I}_p)^{-1/2}(\Sigma + \lambda\mathbf{I}_p)^{1/2}\|_2^4 \leqslant 100 + \exp(-cN)\|\Sigma\|_2^2/\lambda^2$ for some constant $c > 0$.

Proof of Lemma E.1. Adopt the shorthand notation $\Sigma_\lambda = \Sigma + \lambda\mathbf{I}_p, \widehat{\Sigma}_\lambda = \widehat{\Sigma} + \lambda\mathbf{I}_p$. By some basic algebra, we have

$$\begin{aligned}\|(\widehat{\Sigma} + \lambda\mathbf{I}_p)^{-1/2}\Sigma^{1/2}\|_2^2 \leqslant \|(\widehat{\Sigma} + \lambda\mathbf{I}_p)^{-1/2}(\Sigma + \lambda\mathbf{I}_p)^{1/2}\|_2^2 &= \|\Sigma_\lambda^{1/2}\widehat{\Sigma}_\lambda^{-1}\Sigma_\lambda^{1/2}\|_2 \\ &= \|(\mathbf{I}_p - \Sigma_\lambda^{-1/2}(\Sigma - \widehat{\Sigma})\Sigma_\lambda^{-1/2})^{-1}\|_2. \end{aligned} \tag{33}$$

Let $B = \Sigma_\lambda^{-1/2}\Sigma^{1/2}$. Then we have $\|B\|_2 \leqslant 1$ and $\mathrm{tr}(BB^\top) = \sum_{i=1}^p \frac{\mu_i(\Sigma)}{\mu_i(\Sigma)+\lambda} \leqslant N/4$ by assumption. Therefore, by Theorem 4 and 5 in Koltchinskii and Lounici (2017)

$$\begin{aligned}\|\Sigma_\lambda^{-1/2}(\Sigma - \widehat{\Sigma})\Sigma_\lambda^{-1/2}\|_2 &\leqslant \|B\|_2^2 \cdot \max\left\{\sqrt{\frac{\mathrm{tr}(BB^\top)}{N}}, \frac{\mathrm{tr}(BB^\top)}{N}\right\} + c\sqrt{\frac{t}{N}} \cdot \|B\|_2^2 \\ &\leqslant \sqrt{\frac{\mathrm{tr}(BB^\top)}{N}} + c\sqrt{\frac{t}{N}} \leqslant \frac{1}{2} + c\sqrt{\frac{t}{N}}.\end{aligned}$$

with probability at least $1 - e^{-t}$ for any $t \in [1, N]$. Choosing $t = N/c'$ for some sufficiently large constant $c' > 0$ yields $\|\Sigma_\lambda^{-1/2}(\Sigma - \widehat{\Sigma})\Sigma_\lambda^{-1/2}\|_2 \leqslant 2/3$ with probability at least $1 - e^{-\Omega(N)}$. Combining this with Eq. (33) yields the first part of Lemma E.1.

To establish the bound in expectation, we first use Eq. (33) to obtain an always-valid upper bound

$$\|(\widehat{\Sigma} + \lambda\mathbf{I}_p)^{-1/2}(\Sigma + \lambda\mathbf{I}_p)^{1/2}\|_2^2 \leqslant \frac{1}{\mu_{\min}(\mathbf{I}_p - \Sigma_\lambda^{-1/2}\Sigma\Sigma_\lambda^{-1/2})} = \frac{\lambda + \|\Sigma\|_2}{\lambda}.$$

Combining this with the first part of Lemma E.1, we obtain

$$\mathbb{E}\|(\widehat{\Sigma} + \lambda\mathbf{I}_p)^{-1/2}(\Sigma + \lambda\mathbf{I}_p)^{1/2}\|_2^4 \leqslant 100 + \frac{\exp(-cN)}{\lambda^2} \cdot \|\Sigma\|_2^2$$

for some constant $c > 0$. □

In the next three lemmas, we let $(\lambda_i)_{i=1}^d$ denote the eigenvalues of $\mathbf{H}$ in non-increasing order.

Lemma E.2 (Lemma G.1 in Lin et al. (2024))**.** *Let $\mathbf{S} \in \mathbb{R}^{M\times d}$ be a random sketching matrix with i.i.d. entries $\mathbf{S}_{ij} \sim \mathcal{N}(0, 1/M)$.*[5] *Then there exists some absolute constant $c > 1$ such that for any $M \geqslant 1$ and $0 \leqslant k \leqslant M$, with probability at least $1 - e^{-\Omega(M)} - e^{-\Omega(k)}$, we have*

$$\text{for every } j \leqslant M, \quad \left|\tilde{\lambda}_j - \left(\lambda_j + \frac{\sum_{i>k}\lambda_i}{M}\right)\right| \leqslant c\left(\sqrt{\frac{k}{M}}\lambda_j + \lambda_{k+1} + \sqrt{\frac{\sum_{i>k}\lambda_i^2}{M}}\right).$$

Consequently, if $k \leqslant M/c^2$, then

$$\text{for every } j \leqslant M, \quad \left|\tilde{\lambda}_j - \left(\lambda_j + \frac{\sum_{i>k}\lambda_i}{M}\right)\right| \leqslant \frac{1}{2}\left(\lambda_j + \frac{\sum_{i>k}\lambda_i}{M}\right) + c_1\lambda_{k+1},$$

where $c_1 = c + 2c^2$.

[5] d can be $+\infty$.

Lemma E.3 (Tail concentration; Lemma G.2 in Lin et al. (2024) and Lemma 26 in Bartlett et al. (2020)). *Let $\mathbf{S} \in \mathbb{R}^{M\times d}$ be a random sketching matrix with i.i.d. entries $\mathbf{S}_{ij} \sim \mathcal{N}(0, 1/M)$. For any $k \geqslant 0$, with probability at least $1-\delta$, we have*

$$\left\|\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top - \frac{\sum_{i>k}\lambda_i}{M}\cdot\mathbf{I}_M\right\|_2 \lesssim \frac{1}{M}\left(\lambda_{k+1}\left(M+\log\frac{1}{\delta}\right)+\sqrt{\sum_{i>k}\lambda_i^2\left(M+\log\frac{1}{\delta}\right)}\right).$$

In particular, with probability at least $1-e^{-\Omega(M)}$, we have

$$\left\|\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top - \frac{\sum_{i>k}\lambda_i}{M}\cdot\mathbf{I}_M\right\|_2 \lesssim \lambda_{k+1}+\sqrt{\frac{\sum_{i>k}\lambda_i^2}{M}}.$$

Furthermore, the minimum eigenvalue of $\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top$ satisfies

$$\mu_{\min}\left(\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top\right) \gtrsim \lambda_{k+2M}$$

with probability at least $1-e^{-\Omega(M)}$.

Lemma E.4 (Head concentration; Lemma G.3 in Lin et al. (2024)). *Let $\mathbf{S} \in \mathbb{R}^{M\times d}$ be a random sketching matrix with i.i.d. entries $\mathbf{S}_{ij} \sim \mathcal{N}(0, 1/M)$. For any $k \geqslant 1$, with probability at least $1-\delta$, we have*

$$\text{for every } j \leqslant k, \quad \left|\mu_j\left(\mathbf{S}_{0:k}\mathbf{H}_{0:k}\mathbf{S}_{0:k}^\top\right)-\lambda_j\right| \lesssim \sqrt{\frac{k+\log(1/\delta)}{M}}\lambda_j.$$

In particular, with probability at least $1-e^{-\Omega(k)}$,

$$\text{for every } j \leqslant k, \quad \left|\mu_j\left(\mathbf{S}_{0:k}\mathbf{H}_{0:k}\mathbf{S}_{0:k}^\top\right)-\lambda_j\right| \lesssim \sqrt{\frac{k}{M}}\lambda_j.$$

E.2 Concentration bounds under power-law spectrum

Lemma E.5 (Eigenvalues of $\mathbf{SHS}^\top$ under power-law spectrum; Lemma G.4 in Lin et al. (2024)). *Let Assumption 1C hold. There exist some a-dependent constants $c_2 > c_1 > 0$ such that*

$$c_1 j^{-a} \leqslant \mu_j(\mathbf{SHS}^\top) \leqslant c_2 j^{-a}$$

with probability at least $1-e^{-\Omega(M)}$.

Lemma E.6 (Ratio of eigenvalues of $\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top$ under power-law spectrum; Lemma G.5 in Lin et al. (2024)). *Let Assumption 1C hold. There exists some a-dependent constant $c > 0$ such that for any $k \geqslant 0$, the ratio between the $M/2$-th and M-th eigenvalues of $\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top$ satisfies*

$$\frac{\mu_{M/2}(\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top)}{\mu_M(\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top)} \leqslant c$$

with probability at least $1-e^{-\Omega(M)}$.

Lemma E.7 (Bounds on Approx under the source condition; Lemma C.5 in Lin et al. (2024)). *Suppose Assumption 1 is in force. Then with probability at least $1-e^{-\Omega(M)}$ over $\mathbf{S}$,*

$$M^{1-b} \lesssim \mathbb{E}_{\mathbf{w}*}[\mathsf{Approx}] \lesssim M^{1-b}.$$

Here, the hidden constants only depend on (a, b) in Assumption 1.

Lemma E.8 (Eigenvalues of $\widehat{\mathbf{\Sigma}}$ under power-law spectrum). *Suppose $\mathbf{\Sigma} = \mathbf{SHS}^\top$ satisfies $\mu_j(\mathbf{\Sigma}) \eqsim j^{-a}$ for $j \in [M]$. Then for some a-dependent constants $c, c_1, c_2 > 0$, $\widehat{\mathbf{\Sigma}} = \frac{1}{N}\sum_{i=1}^N \mathbf{S}\mathbf{x}_i\mathbf{x}_i^\top\mathbf{S}^\top$ satisfies*

$$c_1 j^{-a} \leqslant \mu_j(\widehat{\mathbf{\Sigma}}) \leqslant c_2 j^{-a} \text{ for all } j \leqslant \min\{M, N/c\}, \quad \text{and}$$
$$\mu_j(\widehat{\mathbf{\Sigma}}) \leqslant c_2 j^{-a} \text{ for all } j \in (\min\{M, N/c\}, \min\{M, N\}]$$

with probability at least $1-e^{-\Omega(N)}$ over the randomness of $(\mathbf{x}_i)_{i=1}^N$ conditioned on $\mathbf{S}$.

Proof of Lemma E.8. Note that $\mathbf{S}\mathbf{X}^\top/\sqrt{N} \overset{d}{=} \mathbf{\Sigma}^{1/2}\mathbf{Z}^\top$, where $\mathbf{Z} \in \mathbb{R}^{M\times N}$ has i.i.d. entries $\mathbf{Z}_{ij} \sim \mathcal{N}(0, 1/N)$ conditioned on $\mathbf{S}$. Thus, $\mu_j(\widehat{\mathbf{\Sigma}}) = \mu_j(\mathbf{Z}\mathbf{\Sigma}\mathbf{Z}^\top)$ for $j \leqslant \min\{M, N\}$. Let $(\widehat{\lambda}_i)_{i=1}^N$ denote the eigenvalues of $\mathbf{Z}\mathbf{\Sigma}\mathbf{Z}^\top$ in non-increasing order. Using Lemma E.2 with $k = N/c$ for some sufficiently large constant c and noting that $\sum_{i>k} i^{-a} \lesssim k^{1-a}$, we have

$$\frac{1}{2}\cdot(j^{-a} + \tilde{c}_1 N^{-a}) - \tilde{c}_2 \cdot N^{-a} \leqslant \widehat{\lambda}_j \leqslant \frac{3}{2}\cdot(j^{-a} + \tilde{c}_1 N^{-a}) + \tilde{c}_2 \cdot N^{-a}$$

for every $j \leqslant \min\{M, N/c\}$ for some constants $\tilde{c}_i, i \in [2]$ with probability at least $1 - e^{-\Omega(N)}$. Therefore, for all $j \leqslant \min\{M, N/\tilde{c}\}$ for some sufficiently large constant $\tilde{c} > 1$, we have

$$\widehat{\lambda}_j \in [\tilde{c}_3 j^{-a}, \tilde{c}_4 j^{-a}]$$

with probability at least $1 - e^{-\Omega(N)}$ for some constants $\tilde{c}_3, \tilde{c}_4 > 0$. For $j \in (\min\{M, N/\tilde{c}\}, \min\{M, N\}]$, by monotonicity of the eigenvalues, we have

$$\widehat{\lambda}_j \leqslant \widehat{\lambda}_{\lfloor \min\{M, N/\tilde{c}\}\rfloor} \leqslant \tilde{c}_4\Big(\Big\lfloor \min\{M, N/\tilde{c}\}\Big\rfloor\Big)^{-a} \leqslant \tilde{c}_5 \min\{M, N\}^{-a} \leqslant \tilde{c}_5 j^{-a}$$

for some sufficiently large constant $\tilde{c}_5 > \tilde{c}_4$ with probability at least $1 - e^{-\Omega(N)}$. □