Risk Bounds of Multi-Pass SGD for Least Squares in the Interpolation Regime

Anonymous Author(s)
Affiliation
Address
email

Abstract

Stochastic gradient descent (SGD) has achieved great success due to its superior performance in both optimization and generalization. Most of existing generalization analyses are made for single-pass SGD, which is a less practical variant compared to the commonly-used multi-pass SGD. Besides, theoretical analyses for multi-pass SGD often concern a worst-case instance in a class of problems, which may be pessimistic to explain the superior generalization ability for some particular problem instance. The goal of this paper is to provide an instance-dependent and algorithm-dependent excess risk bound of multi-pass SGD for least squares in the interpolation regime, which is expressed as a function of the iteration number, stepsize, and data covariance. We show that the excess risk of SGD can be exactly decomposed into the excess risk of GD and a positive fluctuation error, suggesting that SGD always performs worse, instance-wisely, than GD, in generalization. On the other hand, we show that although SGD needs more iterations than GD to achieve the same level of excess risk, it saves the number of stochastic gradient evaluations, and therefore is preferable in terms of computational time.

1 Introduction

Stochastic gradient descent (SGD) is one of the workhorses in modern machine learning due to its efficiency and scalability in training and good ability in generalization to unseen test data. From the optimization perspective, the efficiency of SGD is well understood. For example, to achieve the same level of optimization error, SGD saves the number of gradient computations compared to its deterministic counterpart, i.e., batched *gradient descent* (GD) [7, 8], and therefore saves the total amount of running time. However, the generalization ability (e.g., excess risk bounds) of SGD is far less clear, especially from theoretical perspective.

Single-pass SGD, a less practical SGD variant where each training data is used only once, has been extensively studied in theory. In particular, a series of works establishes excess risk bounds of single-pass SGD for learning general smooth and convex objectives [32, 24, 26] as well as learning least squares [3, 10, 18, 19, 27, 13, 43, 39]. In practice, though, one often runs SGD with *multiple passes* over the training data and outputs the final iterate, which is referred to as *multi-pass SGD* (or simply SGD in the rest of this paper when there is no confusion). Compared to single-pass SGD that has limited number of optimization steps, multi-pass SGD allows the algorithm to perform arbitrary number of optimization steps, which is more powerful in optimizing the empirical risk and thus leads to smaller bias error [28].

Despite the extensive application of multi-pass SGD in practice, there are only a few theoretical techniques being developed to study the generalization of multi-pass SGD. One is based on the *uniform stability* [12, 16], which is defined as the change of the model outputs under a small change in the training data. However, the stability based generalization bound is a worst-case guarantee,

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

which is relatively crude and does not show difference between GD and SGD (See, e.g., Chen et al. [9] showed GD and SGD have the same stability parameter in the convex smooth setting). On the contrary, one easily observes a generalization difference between SGD and GD even in learning the simplest least square problem (see Figure 1). In addition, Lin and Rosasco [22], Pillaud-Vivien et al. [28], Mücke et al. [25] explored the risk bounds for multi-pass SGD using the *operator methods* that are originally developed for analyzing single-pass SGD. Their bounds are sharp in the minimax sense for a class of least square problems that satisfy certain *source condition* (which restricts the norm of the optimal parameter) and *capacity condition* (or effective dimension, which restricts the spectrum of the data covariance matrix). Still, their bounds are uniform for a class of problem instances, and cannot point-wisely adapt to each problem instance (e.g., a least square problem with a particular data covariance matrix). In particular their reults are pessimistic for the benign-overfitting [4] least square instances (see Theorem 4.2 and related discussions).

In this paper, our goal is to establish an algorithm-dependent and problem-dependent excess risk bound of multi-pass SGD for least squares. Our focus is the *interpolation regime* where the training data can be perfectly fitted by a linear interpolator (which holds almost surely when the number of parameter d exceeds the number of training data n). We assume the data has a sub-Gaussian tail [4]. Our main contributions are summarized as follows:

- We show that for any iteration number and stepsize, the excess risk of SGD can be exactly decomposed into the excess risk of GD (with the same stepsize and iteration number) and the so-called *fluctuation error*, which is attributed to the accumulative variance of stochastic gradients in all iterations. This suggests that GD (with optimally tuned hyperparameters) always achieves a smaller excess risk than SGD for least square problems.
- We further establish problem-dependent bounds for the excess risk of GD and the fluctuation error, stated as a function of the eigenspectrum of the data covariance, iteration number, training sample size, and stepsize. Compared to the bounds proved in prior works [22, 28, 25], our bounds allow a wider range of iteration numbers t, and correctly vanishes when $t \to \infty$ in the benign overfitting regime [4]. In contrast, the prior results do not allow $t \to \infty$.
- We develop a new suite of proof techniques for analyzing the excess risk of multi-pass SGD. The key idea is considering the error in its matrix form and how it is updated based on the tensor operators defined by the second-order and fourth-order moments of the empirical data distribution (i.e., sampling with replacement from the training dataset), rather than the operators used in the single-pass SGD analysis that are defined based on the population data distribution [18, 43], together with a sharp characterization on the properties of the operators.

Based on the excess risk upper bounds for SGD and GD, we make the following complexity comparison between SGD and GD: to achieve the same order of excess risk, while SGD may need more iterations than GD, it can have fewer stochastic gradient evaluations than GD. For example, consider the case that the data covariance matrix has a polynomially decaying spectrum with rate $i^{-(1+r)}$, where $r > 0$ is an absolute constant. In order to achieve the same order of excess risk, we have the following comparison in terms of iteration complexity and gradient complexity[1]:

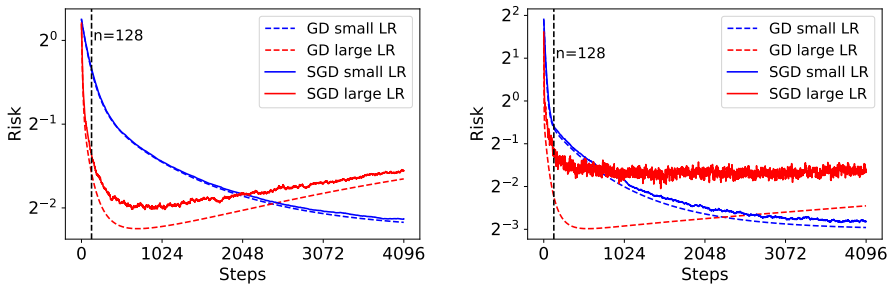
- *Iteration Complexity:* SGD needs to take $\widetilde{\mathcal{O}}(n^{\max\{0.5, \frac{r}{r+1}\}})$ more iterations than GD, with optimally tuned iteration number and stepsize.
- *Gradient Complexity:* SGD needs $\widetilde{\mathcal{O}}(n^{\max\{0.5, \frac{1}{r+1}\}})$ less stochastic gradient evaluations than GD.

Notation. For $n > 0$, we use $\mathrm{poly}(n)$ to define some positive high-degree polynomial functions of n. For two positive-value functions $f(x)$ and $g(x)$ we write $f(x) \lesssim g(x)$ if $f(x) \leq cg(x)$ for some constant $c > 0$, we write $f(x) \gtrsim g(x)$ if $g(x) \lesssim f(x)$, and $f(x) \eqsim g(x)$ if both $f(x) \lesssim g(x)$ and $g(x) \lesssim f(x)$ hold. We use $\widetilde{\mathcal{O}}(\cdot)$ to hide some polylogarithmic factors in the standard big-$\mathcal{O}$ notation. For two matrices $\mathbf{A}$ and $\mathbf{B}$, we denote $\langle \mathbf{A}, \mathbf{B} \rangle = \sum_{i,j} \mathbf{A}_{ij}\mathbf{B}_{ij}$ and $\mathbf{A} \otimes \mathbf{B}$ as their Kronecker product.

2 Related Work

Optimization. Regarding optimization efficiency, the benefit of SGD is well understood [7, 8, 23, 5, 34, 35, 26]. For example, for strongly convex losses (can be relaxed with certain growth conditions), GD has less iteration complexity, but SGD enjoys less gradient complexity [7, 8].

[1] We define the gradient complexity as the number of required stochastic gradient evaluations to achieve a target excess risk, which is closely related to the total computation time.



(a) $\lambda_i = i^{-1}\log^{-2}(i+10)$ (b) $\lambda_i = i^{-2}$

Figure 1: Excess risk comparison between SGD and GD with large and small stepsizes. The true parameter $\mathbf{w}^*$ is randomly drawn from $\mathcal{N}(0, \mathbf{I})$ and the model noise variance $\sigma^2 = 1$. The problem dimension is $d = 256$, and we randomly draw $n = 128$ training data. We consider two data covariance with eigenspectrum $\lambda_i = i^{-1}\log^{-2}(i+10)$ and $\lambda_i = i^{-2}$. For SGD, the reported risk is averaged over 100 repeats of the algorithm's randomness. The large stepsize is $\eta = 0.2$ and the small stepsize is $\eta = 0.02$.

More recently, it is shown that SGD can converge at an exponential rate in the interpolating regime [26, 23, 5, 34, 35], therefore SGD can match the iteration complexity of GD. Nevertheless, all the above results are regrading the optimization performance; our focus in this paper is to study the generalization performance of SGD (and GD).

Risk Bounds for Multi-Pass SGD. The risk bounds of multi-pass SGD are also studied from the operator perspective [30, 22, 28, 25]. The work by Rosasco and Villa [30] focused on *cyclic SGD*, i.e., SGD with multiple passes but fixed sequence on the training data. Their results are limited to small stepsizes ($\gamma = \mathcal{O}(1/n)$), while ours allow constant stepsize. Similar to Lin and Rosasco [22], Pillaud-Vivien et al. [28], Mücke et al. [25], we decompose the population risk of SGD iterates into a risk term caused by batch GD iterates and a fluctuation error term between SGD and GD iterates. But our methods of bounding the fluctuation error are different (see more in Section 5). Moreover, our results are based on different assumptions: Lin and Rosasco [22], Pillaud-Vivien et al. [28], Mücke et al. [25] assumed finiteness on the optimal parameter, and their results only apply to data covariance with a specific type of spectrum (nearly polynomially decaying ones); in contrast, our results assume a Gaussian prior on the optimal parameter (which might not admit a finite norm), and our results are stated as a function of the entire eigenspectrum of the data covariance, thus cover more general data covariance (including those with polynomially decaying spectrum). Lei et al. [21] studied risk bounds for multi-pass SGD with general convex loss. When applied to least square problems, their bounds are cruder than ours.

Uniform Stability. Another approach for characterizing the generalization of multi-pass SGD is through *uniform stability* [16, 9, 20, 42, 6]. There are mainly two differences between this and our approach. First, we directly bound the excess risk of SGD; but the uniform stability can only bound the generalization error, there needs an additional triangle inequality to relate excess risk with generalization error plus optimization error (plus approximation error) — this inequality can easily be loose (consider the algorithmic regularization effects). Secondly, the uniform stability bound is also crude. For example, in the non-strongly convex setting, the uniform stability bound for SGD/GD linearly scales with the total optimization length (i.e., sum of stepsizes), which grows as t [16, 9, 20, 42, 6] (this is minimaxly unavoidable according to Zhang et al. [42], Bassily et al. [6]). Notably, Bassily et al. [6] extended the uniform stability approach to the non-convex and smooth setting. We left such an extension of our method as a future work.

3 Problem Setup

Let $\mathbf{x}$ be a feature vector in a Hilbert space $\mathcal{H}$ (its dimension is denoted by d, which is possibly infinite) and $y \in \mathbb{R}$ be its response, and assume that they jointly follow an unknown population distribution $\mathcal{D}$. In linear regression problems, the *population risk* of a parameter $\mathbf{w}$ is defined by

$$L_{\mathcal{D}}(\mathbf{w}) := \frac{1}{2}\mathbb{E}_{(\mathbf{x},y)\sim D}(\langle \mathbf{x}, \mathbf{w}\rangle - y)^2,$$

and the *excess risk* is defined by

$$\mathcal{E}(\mathbf{w}) := L_{\mathcal{D}}(\mathbf{w}) - \min_{\mathbf{w}} L_{\mathcal{D}}(\mathbf{w}) = \frac{1}{2}\|\mathbf{w} - \mathbf{w}^*\|_{\mathbf{H}}^2, \quad \text{where } \mathbf{H} := \mathbb{E}_{\mathcal{D}}[\mathbf{x}\mathbf{x}^\top], \tag{3.1}$$

where $\mathbf{w}^* = \arg\min_{\mathbf{w}} L_{\mathcal{D}}(\mathbf{w})$ denotes the global minimizer of the population risk. Additionally, following Zou et al. [43, 44], we assume that the data covariance matrix $\mathbf{H}$ is positive definite. In the statistical learning setting, the population distribution $\mathcal{D}$ is unknown, and one is provided with a set of n training samples, $\mathcal{S} = (\mathbf{x}_i, y_i)_{i=1}^n$, that are drawn independently at random from the population distribution. We also use $\mathbf{X} := (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$ and $\mathbf{y} := (y_1, \ldots, y_n)^\top$ to denote the concatenated features and labels, respectively. The linear regression problems aim to find a parameter based on the training set $\mathcal{S}$ that affords a small excess risk.

Multi-Pass SGD. We are interested in solving the linear regression problem using multi-pass *stochastic gradient descent* (SGD)[2] with a constant learning rate. The algorithm generates a sequence of iterates $(\mathbf{w}_t)_{t\geq 1}$ according to the following update rule: the initial iterate is $\mathbf{w}_0 = \mathbf{0}$ (which can be assumed without loss of generality); then at each iteration, an example $(\mathbf{x}_{i_t}, y_{i_t})$ is drawn from $\mathcal{S}$ uniformly at random, and the iterate is updated by

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \cdot \mathbf{x}_{i_t}(\mathbf{x}_{i_t}^\top \mathbf{w}_t - y_{i_t}),$$

where $\eta > 0$ is a constant stepsize (i.e., learning rate).

GD. Another popular algorithm is *gradient descent* (GD). For the clarity of notations, we use $(\widehat{\mathbf{w}}_t)_{t\geq 1}$ to denote the GD iterates, which follow the following updates:

$$\widehat{\mathbf{w}}_{t+1} = \widehat{\mathbf{w}}_t - \eta \cdot \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i(\mathbf{x}_i^\top \widehat{\mathbf{w}}_t - y_i), \quad \widehat{\mathbf{w}}_0 = \mathbf{0},$$

where $\eta > 0$ is a constant stepsize.

Definitions and Assumptions. The eigenvalues of the population data covariance $\mathbf{H}$ is denoted by $(\lambda_i)_{i\geq 1}$, sorted in non-increasing order. Given the training data $(\mathbf{X}, \mathbf{y})$, we define $\boldsymbol{\epsilon} = \mathbf{y} - \mathbf{X}\mathbf{w}^*$ the collection of model noise, $\mathbf{A} = \mathbf{X}\mathbf{X}^\top$ as the Gram matrix, and $\boldsymbol{\Sigma} = n^{-1}\mathbf{X}^\top\mathbf{X}$ as the empirical covariance. Then the *minimum-norm solution* is defined by

$$\widehat{\mathbf{w}} := (\mathbf{X}^\top\mathbf{X})^\dagger \mathbf{X}^\top \mathbf{y} = \mathbf{X}^\top \mathbf{A}^{-1}\mathbf{y}.$$

It is clear that in the interpolation regime, with appropriate stepsizes, both SGD and GD algorithms converge to $\widehat{\mathbf{w}}$ [14, 4].

The assumptions required by our theorems are summarized in below.

Assumption 3.1 *For the linear regression problem:*

- *A The components of $\mathbf{H}^{-1/2}\mathbf{x}$ are independent and 1-subGaussian.*
- *B The response y is generated by $y := \langle \mathbf{w}^*, \mathbf{x}\rangle + \xi$, where $\mathbf{w}^*$ is the ground truth weight vector and ξ is a noise independent of $\mathbf{x}$. Furthermore, the additive noise satisfies $\mathbb{E}[\xi] = 0$, $\mathbb{E}[\xi^2] \leq \sigma^2$.*
- *C The ground truth $\mathbf{w}^*$ follows a Gaussian prior $\mathcal{N}(\mathbf{0},\ \omega^2 \cdot \mathbf{I})$, where ω^2 is a constant.*
- *D The minimum-norm solution $\widehat{\mathbf{w}}$ linearly interpolates all training data, i.e., $y_i = \widehat{\mathbf{w}}^\top \mathbf{x}_i$ for $i \in [n]$.*

Assumptions 3.1A and B are standard for analyzing overparameterized linear regression problem in the benign overfitting regime [4, 33]. Note that Assumption 3.1A is widely made in the analysis of high-dimensional least squares estimations [11, 38, 17]. However, Assumption 3.1A is not standard for analyzing SGD [3, 22, 28, 18, 43]. We conjecture Assumption 3.1A can be relaxed and leave this as a future work. Moreover, assumption 3.1C is also widely adopted in analyzing least square problems (see, e.g., Ali et al. [2], Dobriban et al. [11], Xu and Hsu [40]). Finally, Assumption 3.1D holds almost surely when $d > n$, i.e., the number of parameter exceeds the number of data.

In the following, the presented risk bounds will hold (i) with high-probability with respect to the randomness of sampling feature vectors $\mathbf{X}$, and (ii) in expectation with respect to the randomness of multi-pass SGD algorithm, the randomness of sampling additive noise $\boldsymbol{\epsilon}$ and the randomness of the true parameter $\mathbf{w}^*$ as a prior. For these purpose, we will use $\mathbb{E}_{i_t}, \mathbb{E}_{\mathrm{SGD}}, \mathbb{E}_{\mathbf{w}^*}$ to refer to taking expectation with respect to the randomness of sampling data (from the training set) at the t-th iteration, the randomness of the entire SGD algorithm (i.e., sampling data at each iteration, $i_1, \ldots, i_t, \ldots$) and the prior distribution of $\mathbf{w}^*$, respectively.

[2] We focus on *SGD with replacement* in this paper. Extending our results to SGD without replacement is an important yet challenging future direction (see more discussions in Section 6).

4 Main Results

Our first theorem shows that, under the same stepsize and number of iterates, SGD always generalizes worse than GD.

Theorem 4.1 (Risk decomposition) *Suppose that Assumption 3.1D holds. Then the excess risk of SGD can be decomposed by*

$$\mathbb{E}_{\mathrm{SGD}}\big[\mathcal{E}(\mathbf{w}_t)\big] = \mathcal{E}(\widehat{\mathbf{w}}_t) + \mathrm{FluctuationError}(\mathbf{w}_t).$$

Moreover, the fluctuation error is always non-negative.

A Risk Comparison. Theorem 4.1 shows that, in the interpolation regime, SGD affords a strictly larger excess risk than GD, given the same hyperparameters (stepsize η and number of iterates t). Therefore, despite of a possibly higher computational cost, the optimally tuned GD *dominates* the optimally tuned SGD in terms of the generalization performance. This observation is verified empirically by experiments in Figure 1.

Theorem 4.1 relates the risk of SGD iterates to that of GD iterates. This idea has appeared in earlier literature [22, 28, 25]. However, their decomposition is obtained via Young's inequality (see, e.g., Eq. (13) in Appendix A of Mücke et al. [25]), and is therefore stated as an upper bound on the SGD risk.

Our next theorem is to characterize the fluctuation error of SGD (with respect to GD).

Theorem 4.2 (Fluctuation error bound) *Suppose that Assumptions 3.1A, B and D all hold. Then for every $n \geq 1$, $t \geq 1$ and $\eta \leq c/\operatorname{tr}(\mathbf{H})$ for some absolute constant c, with probability at least $1 - 1/\mathrm{poly}(n)$, it holds that*

$$\mathrm{FluctuationError}(\mathbf{w}_t) \lesssim \left[\log(t) \cdot \left(\frac{\operatorname{tr}(\mathbf{H})\log(n)}{t} + \frac{k^\dagger \log^{5/2}(n)}{n^{1/2}t}\right) + \frac{\log^{5/2}(n)\eta}{n^{1/2}} \cdot \sum_{i>k^\dagger} \lambda_i\right] \cdot \min\left\{\|\widehat{\mathbf{w}}\|_2^2,\ t\eta \cdot \|\widehat{\mathbf{w}}\|_{\mathbf{\Sigma}}^2\right\},$$

where $k^\dagger \geq 0$ is an arbitrary index (can be infinity).

We first explain the factor $\min\left\{\|\widehat{\mathbf{w}}\|_2^2,\ t\eta \cdot \|\widehat{\mathbf{w}}\|_{\mathbf{\Sigma}}^2\right\}$ in our bound. First of all, when the interpolator $\widehat{\mathbf{w}}$ has a small ℓ_2-norm, the quantity is automatically small. Furthermore, $\|\widehat{\mathbf{w}}\|_{\mathbf{\Sigma}}^2 \lesssim \omega^2 \lesssim 1$ easily holds under mild assumptions on $\mathbf{w}^*$, e.g., Assumption 3.1C. Then, for finite t one can bound the factor with $\min\left\{\|\widehat{\mathbf{w}}\|_2^2,\ t\eta \cdot \|\widehat{\mathbf{w}}\|_{\mathbf{\Sigma}}^2\right\} \lesssim \omega^2\eta t$. More interestingly, for SGD with constant stepsize and infinite optimization steps ($t \to \infty$), our risk bound can still vanish, while all risk bounds in prior works [22, 28, 25] are vacuous. To see this, one can consider a sequence of $k^\dagger$, e.g., $k^\dagger = \sqrt{t}$, then it is clear that $\log(t)k^\dagger(t)/t, \sum_{i>k^\dagger(t)} \lambda_i \to 0$ when $t \to \infty$, so the fluctuation error vanishes.

To complement the above results, we provide the following risk bound for GD. We emphasize that any risk bound for GD can be plugged into Theorems 4.1 and 4.2 to obtain a risk bound for SGD.

Theorem 4.3 (GD risk) *Suppose that Assumptions 3.1A, B and C all hold. Then for every $n \geq 1$, $t \geq 1$ and $\eta < 1/\|\mathbf{H}\|_2$, with probability at least $1 - 1/\mathrm{poly}(n)$, it holds that*

$$\mathbb{E}_{\mathbf{w}^*,\boldsymbol{\epsilon}}[\mathcal{E}(\widehat{\mathbf{w}}_t)] \lesssim \omega^2 \cdot \left(\frac{\widetilde{\lambda}^2}{n^2} \cdot \sum_{i \leq k^*} \frac{1}{\lambda_i} + \sum_{i>k^*} \lambda_i\right) + \sigma^2 \cdot \left(\frac{k^*}{n} + \frac{n}{\widetilde{\lambda}^2} \sum_{i>k^*} \lambda_i^2\right),$$

where $k^* := \min\{k : n\lambda_{k+1} \leq \frac{n}{\eta t} + \sum_{i>k} \lambda_i\}$ and $\widetilde{\lambda} := \frac{n}{\eta t} + \sum_{i>k^*} \lambda_i$.

The bound presented in Theorem 4.3 is comparable to that for ridge regression established by Tsigler and Bartlett [33] and will be much better than the bound of single-pass SGD when the signal-to-noise ratio is large [44, Theorem 5.1], e.g., $\omega^2 \gg \sigma^2$. In fact, Theorem 4.3 is proved via a reduction to ridge regression results. In particular, the quantity $n/(\eta t)$ for GD is an analogy to the regularization parameter λ for ridge regression [41, 29, 37, 2]. As a final remark, the assumption that $\mathbf{w}^*$ follows a Gaussian prior is the main concealing in Theorem 4.3 (which is not required by Tsigler and Bartlett [33]). The Gaussian prior on $\mathbf{w}^*$ is known to allow a connection between early stopped GD with ridge regression [2]. We conjecture that this assumption is not necessary and potentially removable.

Combining Theorems 4.1, 4.2 and 4.3, we obtain the following risk bound for multi-pass SGD:

Corollary 4.4 *Suppose that Assumptions 3.1A, B, C and D all hold. Then with probability at least* $1 - 1/\mathrm{poly}(n)$*, it holds that*

$$\mathbb{E}_{\mathrm{SGD},\mathbf{w}^*,\boldsymbol{\epsilon}}\big[\mathcal{E}(\mathbf{w}_t)\big] \lesssim \omega^2 \cdot \left(\frac{\widetilde{\lambda}^2}{n^2} \cdot \sum_{i \le k^*} \frac{1}{\lambda_i} + \sum_{i > k^*} \lambda_i\right) + \sigma^2 \cdot \left(\frac{k^*}{n} + \frac{n}{\widetilde{\lambda}^2} \sum_{i > k^*} \lambda_i^2\right)$$
$$+ \eta \cdot \left[\log(t) \cdot \left(\mathrm{tr}(\mathbf{H}) \log(n) + \frac{k^\dagger \log^{5/2}(n)}{n^{1/2}}\right) + \frac{\log^{5/2}(n) t\eta}{n^{1/2}} \cdot \sum_{i > k^\dagger} \lambda_i\right)\Bigg]$$
$$\cdot \min\left\{(t\eta)^{-1} \cdot \big(n\omega^2 + \sigma^2 \,\mathrm{tr}(\mathbf{A}^{-1})\big), \omega^2 \,\mathrm{tr}(\mathbf{H}) + \sigma^2\right\},$$

where $k^\dagger$ *is an arbitrary index,* $k^* := \min\{k : n\lambda_{k+1} \le \frac{n}{\eta t} + \sum_{i>k} \lambda_i\}$ *and* $\widetilde{\lambda} := \frac{n}{\eta t} + \sum_{i > k^*} \lambda_i$.

Comparison with Existing Results. We now discuss relationships between our bound and existing ones for multi-pass SGD [22, 28, 25]. First, we highlight that our bound is *problem-dependent* in the sense that the bound is stated as a function of the spectrum of data covariance; in contrast, existing papers only provide a minimax analysis for multi-pass SGD. Secondly, we rely on a different set of assumptions from the aforementioned papers. In particular, Pillaud-Vivien et al. [28] requires a *source condition* on the data covariance (e.g. $\|\mathbf{H}^{1/2-r}\mathbf{w}^*\|_2 < \infty$ for some constant $r \ge 0$), and Lin and Rosasco [22], Mücke et al. [25] require an *effective dimension* to be small, but our results are more general regarding the data covariance. Moreover, we assume $\mathbf{w}^*$ follows a Gaussian prior (Assumption 3.1C), which is also not directly comparable to the *source condition* in existing works.

Bartlett et al. [4] showed that OLS generalizes in the so called *benign overfitting* regime. Since when $t \to \infty$, SGD (with constant stepsize) converges to OLS, it would be interesting to compare the SGD solution with OLS in such a regime. We will do so with the following Corollary 4.5.

Corollary 4.5 *Suppose that Assumptions 3.1A, B, C and D all hold. Assume the spectrum of* $\mathbf{H}$ *satisfies* $\lambda_i = i^{-1}\log(i+1)^{-\beta}$ *for some absolute constant* $\beta > 1$*, then with probability at least* $1 - 1/\mathrm{poly}(n)$*, there exists a choice of* t *and* η *such that*

$$\mathbb{E}_{\mathrm{SGD},\mathbf{w}^*,\boldsymbol{\epsilon}}[\mathcal{E}(\mathbf{w}_t)] \lesssim \omega^2 \cdot \log(n)^{1-\beta} + \sigma^2 \log(n)^{-\beta}.$$

Besides, for any fixed stepsize η*, we have*

$$\lim_{t\to\infty} \mathbb{E}_{\mathrm{SGD},\mathbf{w}^*,\boldsymbol{\epsilon}}[\mathcal{E}(\mathbf{w}_t)] \lesssim \omega^2 \cdot \log(n)^{1-\beta} + \sigma^2 \log(n)^{-1}.$$

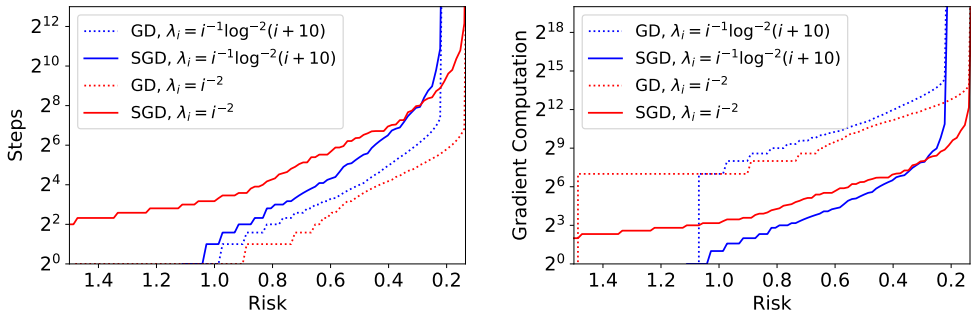
As a sanity check, our bound for $t \to \infty$ matches the upper and lower bounds on the excess risk of OLS (which can be obtained by setting $\lambda = 0$ in Theorem 1 and Lemmas 2 & 3 in Tsigler and Bartlett [33]). Moreover, Corollary 4.5 suggests that the excess risk achieved by multipass SGD is *always no worse than* that of OLS, and could be *strictly smaller than* that of OLS when $\beta > 0$.This demonstrates the benefit of multi-pass SGD over OLS.

The following corollary characterizes the risk of multi-pass SGD for data covariance with a polynomially decaying spectrum.

Corollary 4.6 *Suppose that Assumptions 3.1A, B, C and D all hold. Assume the spectrum of* $\mathbf{H}$ *decays polynomially, i.e.,* $\lambda_i = i^{-1-r}$ *for some absolute constant* $r > 0$*, then with probability at least* $1 - 1/\mathrm{poly}(n)$*, it holds that*

$$\mathbb{E}_{\mathbf{w}^*,\boldsymbol{\epsilon}}[\mathcal{E}(\widehat{\mathbf{w}}_t)] \lesssim \omega^2 \cdot (t\eta)^{-r/(r+1)} + \sigma^2 \cdot \frac{(t\eta)^{1/(r+1)}}{n},$$
$$\mathbb{E}_{\mathrm{SGD},\mathbf{w}^*,\boldsymbol{\epsilon}}[\mathcal{E}(\mathbf{w}_t)] \lesssim \omega^2 \cdot (t\eta)^{-r/(r+1)} + \sigma^2 \cdot \frac{(t\eta)^{1/(r+1)}}{n}$$
$$+ (\omega^2 + \sigma^2) \cdot \eta \cdot \log(t) \cdot \left[\log(n) + \frac{\log^{5/2}(n)}{n^{1/2}} \cdot (t\eta)^{1/(r+1)}\right].$$

Corollary 4.6 provides concrete excess risk bounds for SGD and GD, based on which we can make a comparison between SGD and GD in terms of their iteration and gradient complexities. For simplicity, in the following discussion, we assume that $\omega^2 \eqsim \sigma^2 \eqsim 1$. Then choosing $t\eta \eqsim n$ minimizes the



(a) Iteration Complexity vs. Risk

(b) Gradient Complexity vs. Risk

Figure 2: Iteration and gradient complexity comparison between SGD and GD. The curves report the minimum number of steps/gradients for each algorithm (with an optimally tuned stepsize) to achieve a targeted risk. Experiment setup is the same as that in Figure 1.

upper bound for GD risk and yields the $O(n^{-r/(r+1)})$ rate. Here GD can employ a constant stepsize. Similarly, SGD can match the GD's rate, $O(n^{-r/(r+1)})$, by setting $t\eta \eqsim n$ and

$$\eta \lesssim \log^{-1}(t) \cdot \min\{\log^{-1}(n) \cdot n^{-\frac{r}{r+1}},\ \log^{-\frac{5}{2}}(n) \cdot n^{-\frac{1}{2}})\}. \tag{4.1}$$

The above implies that SGD (fixed stepsize, last iterate) can only cooperate with small stepsize.

Iteration Complexity. We first compare GD and SGD in terms of the iteration complexity. To reach the optimal rate, GD can employ a constant stepsize and set the number of iterates to be $t \eqsim n$. However, in order to shelve the fluctuation error, the stepsize of SGD cannot be large, as required by (4.1). More precisely, in order to match the optimal rate, SGD needs to use a small stepsize, $\eta \eqsim n/t$, with a large number of iterates,

$$t \eqsim \begin{cases} \log(n) \cdot n^{1+\frac{r}{r+1}} = \widetilde{\mathcal{O}}(n^{1+\frac{r}{r+1}}), & r > 1; \\ \log^{3.5}(n) \cdot n^{1.5} = \widetilde{\mathcal{O}}(n^{1.5}), & r \le 1. \end{cases}$$

It can be seen that the iteration complexity of SGD is much worse than that of GD. This result is empirically verified by Figure 2 (a).

Gradient Complexity. We next compare GD and SGD in terms of the gradient complexity. Recall that for each iterate, GD computes n gradients but SGD only computes 1 gradient. Therefore, to reach the optimal rate, the total number of gradient computed by GD needs to be $\Theta(n^2)$, but that computed by SGD is only $\widetilde{\mathcal{O}}(n^{\max\{(2r+1)/(r+1),1.5\}})$. Thus, the gradient complexity of SGD is better than that of GD by a factor of $\widetilde{\mathcal{O}}(n^{\min\{0.5,1/(r+1)\}})$. This result is empirically verified by Figure 2 (b).

5 Overview of the Proof Technique

In this section, we will provide an overview of our proof technique and sketch the proof of Theorems 4.1 and 4.2. The remaining proof is deferred to Appendix.

Our proof technique is inspired by the operator methods for analyzing single-pass SGD [3, 10, 18, 19, 27, 13, 43, 39]. In particular, they track an error *matrix*, $(\mathbf{w}_t - \mathbf{w}^*) \otimes (\mathbf{w}_t - \mathbf{w}^*)$ that keeps richer information than the error norm $\|\mathbf{w}_t - \mathbf{w}^*\|_2^2$. For single-pass SGD where each data is used only once, the resulted iterates enjoy a simple dependence on history that allows an easy calculation of the expected error matrix (with respect to the randomness of data generation). However for multi-pass SGD, a data might be used multiple times, which prevents us from tracking the expected error matrix directly. Instead, a trackable analogy to the error matrix is the *empirical error matrix*, $(\mathbf{w}_t - \widehat{\mathbf{w}}) \otimes (\mathbf{w}_t - \widehat{\mathbf{w}})$ where $\widehat{\mathbf{w}}$ is the minimum norm interpolator. More precisely, note that

$$\mathbf{w}_{t+1} - \widehat{\mathbf{w}} = \mathbf{w}_t - \widehat{\mathbf{w}} - \eta \cdot (\mathbf{x}_{i_t}\mathbf{x}_{i_t}^\top \mathbf{w}_t - \mathbf{x}_{i_t}\mathbf{x}_{i_t}^\top \widehat{\mathbf{w}}) = (\mathbf{I} - \eta \mathbf{x}_{i_t}\mathbf{x}_{i_t}^\top)(\mathbf{w}_t - \widehat{\mathbf{w}}). \tag{5.1}$$

Therefore the expected (over the algorithm's randomness) empirical error matrix updates as follows:

$$\text{let} \quad \mathbf{E}_t := \mathbb{E}_{\text{SGD}}\big[(\mathbf{w}_t - \widehat{\mathbf{w}})(\mathbf{w}_t - \widehat{\mathbf{w}})^\top\big], \text{ then } \mathbf{E}_{t+1} = \mathbb{E}_{i_t}\big[(\mathbf{I} - \eta \mathbf{x}_{i_t}\mathbf{x}_{i_t}^\top)\mathbf{E}_t(\mathbf{I} - \eta \mathbf{x}_{i_t}\mathbf{x}_{i_t}^\top)\big]$$

Let $\mathbf{\Sigma} := \frac{1}{n}\mathbf{X}^\top\mathbf{X}$ be the empirical covariance matrix. We then follow the operator method [43] to define the following operators on symmetric matrices (e.g., $\mathbf{J}$):

$$\mathcal{G} \circ \mathbf{J} := (\mathbf{I} - \eta\mathbf{\Sigma})\mathbf{J}(\mathbf{I} - \eta\mathbf{\Sigma}), \;\; \mathcal{M} \circ \mathbf{J} := \mathbb{E}_{\mathrm{SGD}}[\mathbf{x}_{i_t}\mathbf{x}_{i_t}^\top\mathbf{J}\mathbf{x}_{i_t}\mathbf{x}_{i_t}^\top], \;\; \widetilde{\mathcal{M}} \circ \mathbf{J} := \mathbf{\Sigma}\mathbf{J}\mathbf{\Sigma}.$$

Based on these operators, we can obtain a close form update rule for $\mathbf{E}_t$:

$$\mathbf{E}_t = \mathcal{G} \circ \mathbf{E}_{t-1} + \eta^2 \cdot (\mathcal{M} - \widetilde{\mathcal{M}}) \circ \mathbf{E}_{t-1} = \underbrace{\mathcal{G}^t \circ \mathbf{E}_0}_{\mathbf{\Theta}_1} + \underbrace{\eta^2 \cdot \textstyle\sum_{k=0}^{t-1} \mathcal{G}^{t-1-k} \circ (\mathcal{M} - \widetilde{\mathcal{M}}) \circ \mathbf{E}_k}_{\mathbf{\Theta}_2}. \quad (5.2)$$

Here the first term $\mathbf{\Theta}_1 := (\mathbf{I} - \eta\mathbf{\Sigma})^t\mathbf{E}_0(\mathbf{I} - \eta\mathbf{\Sigma})^t = (\widehat{\mathbf{w}}_t - \widehat{\mathbf{w}})(\widehat{\mathbf{w}}_t - \widehat{\mathbf{w}})^\top$ is exactly the error matrix caused by GD iterates (with stepsize η and iteration number t), and the second term $\mathbf{\Theta}_2$ is a *fluctuation matrix* that captures the deviation of $\mathbf{w}_t$ with respect to a corresponding GD iterate $\widehat{\mathbf{w}}_t$. We remark that the expected error matrix $\mathbf{E}_t$ contains all information of $\mathbf{w}_t$.

Risk Decomposition (Theorem 4.1). The following fact is clear from the update rule (5.1).

Fact 5.1 *The GD iterates satisfy* $\widehat{\mathbf{w}}_{t+1} - \widehat{\mathbf{w}} = (\mathbf{I} - \eta\mathbf{\Sigma})(\widehat{\mathbf{w}}_t - \widehat{\mathbf{w}})$ *and* $\mathbb{E}_{\mathrm{SGD}}[\mathbf{w}_t - \widehat{\mathbf{w}}] = \widehat{\mathbf{w}}_t - \widehat{\mathbf{w}}$.

Based on Fact 5.1 and (5.2), we have

$$\begin{aligned}
&\mathbb{E}_{\mathrm{SGD}}[(\mathbf{w}_t - \mathbf{w}^*)(\mathbf{w}_t - \mathbf{w}^*)^\top] \\
&= \mathbf{E}_t + (\widehat{\mathbf{w}} - \mathbf{w}^*)(\widehat{\mathbf{w}}_t - \widehat{\mathbf{w}})^\top + (\widehat{\mathbf{w}}_t - \widehat{\mathbf{w}})(\widehat{\mathbf{w}} - \mathbf{w}^*)^\top + (\widehat{\mathbf{w}} - \mathbf{w}^*)(\widehat{\mathbf{w}} - \mathbf{w}^*)^\top \\
&= \mathbf{\Theta}_1 + (\widehat{\mathbf{w}} - \mathbf{w}^*)(\widehat{\mathbf{w}}_t - \widehat{\mathbf{w}})^\top + (\widehat{\mathbf{w}}_t - \widehat{\mathbf{w}})(\widehat{\mathbf{w}} - \mathbf{w}^*)^\top + (\widehat{\mathbf{w}} - \mathbf{w}^*)(\widehat{\mathbf{w}} - \mathbf{w}^*)^\top + \mathbf{\Theta}_2 \\
&= (\widehat{\mathbf{w}}_t - \mathbf{w}^*)(\widehat{\mathbf{w}}_t - \mathbf{w}^*)^\top + \mathbf{\Theta}_2,
\end{aligned}$$

where $\mathbf{\Theta}_1$ and $\mathbf{\Theta}_2$ are defined in (5.2) and the last equality is due to $\mathbf{\Theta}_1 = (\widehat{\mathbf{w}}_t - \widehat{\mathbf{w}})(\widehat{\mathbf{w}}_t - \widehat{\mathbf{w}})^\top$. Also note that

$$\mathbb{E}_{\mathrm{SGD}}[\mathcal{E}(\mathbf{w}_t)] = \frac{1}{2}\mathbb{E}_{\mathrm{SGD}}\left[\|\mathbf{w}_t - \mathbf{w}^*\|_{\mathbf{H}}^2\right] = \frac{1}{2}\left\langle \mathbb{E}_{\mathrm{SGD}}[(\mathbf{w}_t - \mathbf{w}^*)(\mathbf{w}_t - \mathbf{w}^*)^\top], \mathbf{H}\right\rangle.$$

Combining these two inequalities proves Theorem 4.1:

$$\mathbb{E}_{\mathrm{SGD}}[\mathcal{E}(\mathbf{w}_t)] = \underbrace{\tfrac{1}{2}\|\widehat{\mathbf{w}}_t - \mathbf{w}^*\|_{\mathbf{H}}^2}_{\text{GD error}} + \underbrace{\tfrac{\eta^2}{2} \cdot \textstyle\sum_{k=0}^{t-1} \left\langle \mathcal{G}^{t-1-k} \circ (\mathcal{M} - \widetilde{\mathcal{M}}) \circ \mathbf{E}_k, \mathbf{H}\right\rangle}_{\text{Fluctuation error}}. \quad (5.3)$$

Finally, the fluctuation error is non-negative because both $\mathcal{G}$ and $\mathcal{M} - \widetilde{\mathcal{M}}$ are PSD mappings.

Bounding the Fluctuation Error (Theorem 4.2). There are several challenges in the analysis of fluctuation error: (1) it is difficult to characterize the matrix $(\mathcal{M} - \widetilde{\mathcal{M}}) \circ \mathbf{E}_k$ since the matrix $\mathbf{E}_k$ is unknown; (2) the operator $\mathcal{G}$ involves an exponential decaying term with respect to the empirical covariance matrix $\mathbf{\Sigma}$, which does not commute with the population covariance matrix $\mathbf{H}$.

To address the first problem, we note that the operators $\widetilde{\mathcal{M}}$, $\mathcal{G}$, $\mathcal{M}$ are PSD mappings and enjoy commutative property and then obtain the following result:

$$\mathrm{FluctuationError} \le \tfrac{\eta^2}{2} \cdot \textstyle\sum_{k=0}^{t-1} \langle \mathcal{M} \circ \mathcal{G}^{t-1-k} \circ \mathbf{H}, \mathbf{E}_k\rangle. \quad (5.4)$$

Now, the input of the operator $\mathcal{M} \circ \mathcal{G}^{t-1-k}$ will not be an unknown matrix but a fixed one (i.e., $\mathbf{H}$), and the remaining effort will be focusing on characterizing $\mathcal{M} \circ \mathcal{G}^k \circ \mathbf{H}$. Applying the definitions of $\mathcal{M}$ and $\mathcal{G}$ implies

$$\mathcal{M} \circ \mathcal{G}^k \circ \mathbf{H} = \mathbb{E}_i\left[\mathbf{x}_i\mathbf{x}_i^\top(\mathbf{I} - \eta\mathbf{\Sigma})^k\mathbf{H}(\mathbf{I} - \eta\mathbf{\Sigma})^k\mathbf{x}_i\mathbf{x}_i^\top\right]. \quad (5.5)$$

Then our idea is to first prove an uniform upper bound on the quantity $\mathbf{x}_i^\top(\mathbf{I} - \eta\mathbf{\Sigma})^k\mathbf{H}(\mathbf{I} - \eta\mathbf{\Sigma})^k\mathbf{x}_i$ for all $i \in [n]$ (e.g., denoted as $U(k, \eta, n)$), then it can be naturally obtained that

$$\mathcal{M} \circ \mathcal{G}^k \circ \mathbf{H} \preceq U(k, \eta, n) \cdot \mathbb{E}_i[\mathbf{x}_i\mathbf{x}_i^\top] = U(k, \eta, n) \cdot \mathbf{\Sigma}, \quad (5.6)$$

then we will only need to characterize the inner product $\langle \mathbf{E}_k, \mathbf{\Sigma}\rangle$ in (5.4), which can be understood as the optimization error at the k-th iteration.

In order to precisely characterize $U(k, \eta, n)$, we encounter the second problem that the population covariance $\mathbf{H}$ and empirical covariance $\mathbf{\Sigma}$ are not commute, thus the exponential decaying term $(\mathbf{I} - \eta\mathbf{\Sigma})^k$ will not be able to fully decrease $\mathbf{H}$ since some components of $\mathbf{H}$ may lie in the small eigenvalue directions of $\mathbf{\Sigma}$. Therefore, we consider the following decomposition

$$\mathbf{x}_i^\top (\mathbf{I} - \eta\mathbf{\Sigma})^k \mathbf{H} (\mathbf{I} - \eta\mathbf{\Sigma})^k \mathbf{x}_i = \underbrace{\mathbf{x}_i^\top (\mathbf{I} - \eta\mathbf{\Sigma})^k \mathbf{\Sigma} (\mathbf{I} - \eta\mathbf{\Sigma})^k \mathbf{x}_i} + \underbrace{\mathbf{x}_i^\top (\mathbf{I} - \eta\mathbf{\Sigma})^k (\mathbf{H} - \mathbf{\Sigma}) (\mathbf{I} - \eta\mathbf{\Sigma})^k \mathbf{x}_i}.$$

Then for Θ_1, it can be seen that the decaying term $(\mathbf{I} - \eta\mathbf{\Sigma})^k$ is commute with $\mathbf{\Sigma}$ thus can successfully make it decrease. For Θ_2, we will view the difference $\mathbf{H} - \mathbf{\Sigma}$ as the component of $\mathbf{H}$ that cannot be effectively decreased by $(\mathbf{I} - \eta\mathbf{\Sigma})^k$, which will be small as n increases. More specifically, we can get the following upper bound on Θ_1.

Lemma 5.2 *If the stepsize satisfies $\gamma \leq c/\operatorname{tr}(\mathbf{H})$ for some small absolute constant c, then with probability at least $1 - 1/\mathrm{poly}(n)$, it holds that $\Theta_1 \lesssim \operatorname{tr}(\mathbf{H}) \cdot \log(n) \cdot \min\left\{\frac{1}{(k+1)\eta},\ \|\mathbf{H}\|_2\right\}$.*

For Θ_2, we will rewrite $\mathbf{x}_i$ as $\mathbf{e}_i^\top \mathbf{X}$ where $\mathbf{e}_i \in \mathbb{R}^n$ and $\mathbf{X} \in \mathbb{R}^{n \times d}$, then

$$\Theta_2 = \mathbf{e}_i^\top \mathbf{X} (\mathbf{I} - \eta\mathbf{\Sigma})^k (\mathbf{H} - \mathbf{\Sigma}) (\mathbf{I} - \eta\mathbf{\Sigma})^k \mathbf{X}^\top \mathbf{e}_i \leq \|\mathbf{e}_i^\top \mathbf{X} (\mathbf{I} - \eta\mathbf{\Sigma})^k\|_2^2 \cdot \|\mathbf{H} - \mathbf{\Sigma}\|_2. \tag{5.7}$$

Then since $\mathbf{X}$ and $\mathbf{\Sigma}$ have the same column eigenspectrum, we can fully unleash the decaying power of the term $(\mathbf{I} - \eta\mathbf{\Sigma})^k$ on $\mathbf{X}$. Further note the that the row space of $\mathbf{X}$ is uniform distributed (corresponding to the index of training data), which is independent of $\mathbf{e}_i$. This implies that we can adopt standard concentration arguments with covering on n fixed vectors $\{\mathbf{e}_i\}_{i=1}^n$ to prove a sharp high probability upper bound (compared to the naive worst-case upper bound). Consequently, we state the upper bound on Θ_2 in the following lemma.

Lemma 5.3 *For every $i \in [n]$ and $k^* \in [d]$, it holds with probability at least $1 - 1/\mathrm{poly}(n)$ that $\Theta_2 \lesssim \frac{\log^{5/2}(n)}{n^{1/2}} \cdot \left(\frac{k^*}{(k+1)\eta} + \sum_{i > k^*} \lambda_i\right)$.*

6 Conclusion and Discussion

In this paper, we establish an instance-dependent excess risk bound of multi-pass SGD for interpolating least square problems. The key takeaways include: (1) the excess risk of SGD is *always* worse than that of GD, given the same setup of stepsize and iteration number; (2) in order to achieve the same level of excess risk, SGD requires more iterations than GD; and (3) however, the gradient complexity of SGD can be better than that of GD. The proposed technique for analyzing multi-pass SGD could be of broader interest.

Several interesting problems are left for future exploration:

A problem-dependent excess risk lower bound could be useful to help understand the sharpness of our excess risk upper bound for multi-pass SGD and establish a clear separation between SGD and GD in terms of iteration and gradient complexity. The challenge here is mainly from the fact that the empirical covariance matrix $\mathbf{\Sigma}$ does not commute with the population covariance matrix $\mathbf{H}$. In particular, one needs to develop an even sharper characterization on the quantity $\mathcal{M} \circ \mathcal{G}^k \circ \mathbf{H}$ (see (5.5)); more precisely, a sharp lower bound on $\mathbf{x}_i^\top (\mathbf{I} - \eta\mathbf{\Sigma})^k \mathbf{H} (\mathbf{I} - \eta\mathbf{\Sigma})^k \mathbf{x}_i$ is required.

SGD with decaying stepsizes could potentially improve the generalization performance of SGD with a constant stepsize. In this regard, most of our analysis can be extended to SGD/GD with decaying stepsizes. For example, Theorem 4.1 directly holds even for varying stepsizes, and Theorem 4.3 also holds under a small modification, i.e., changing $t\eta$ to $\sum_{k=0}^{t-1} \eta_k$. However, our bounds on the fluctuation error may be more subtle to adapt. We conjecture that our analysis idea can still be applied, but the detailed calculations will depend on the particular stepsize scheduler of interests.

Multi-pass SGD without replacement is a more practical SGD variant than the multi-pass SGD with replacement studied in this work. The key difference is that, the former does not pass training data independently (since each data must be used for equal times). In terms of optimization complexity, it has already been demonstrated in theory that multi-pass SGD without replacement (e.g., SGD with single shuffle or random shuffle) outperforms multi-pass SGD with replacement [15, 31, 1]. However, in terms of generalization, whether or not multi-pass SGD without replacement can outperform multi-pass SGD with replacement is still an open problem, as there lacks a sharp excess risk analysis for multi-pass SGD without replacement. The techniques presented in this paper can shed light on this direction.

References

[1] Kwangjun Ahn, Chulhee Yun, and Suvrit Sra. Sgd with shuffling: optimal rates without component convexity and large epoch requirements. *Advances in Neural Information Processing Systems*, 33:17526–17535, 2020.

[2] Alnur Ali, J Zico Kolter, and Ryan J Tibshirani. A continuous-time view of early stopping for least squares regression. In *The 22nd international conference on artificial intelligence and statistics*, pages 1370–1378. PMLR, 2019.

[3] Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $o(1/n)$. *Advances in neural information processing systems*, 26:773–781, 2013.

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

[5] Raef Bassily, Mikhail Belkin, and Siyuan Ma. On exponential convergence of sgd in non-convex over-parametrized learning. *arXiv preprint arXiv:1811.02564*, 2018.

[6] Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. *Advances in Neural Information Processing Systems*, 33:4381–4391, 2020.

[7] Léon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. *Advances in neural information processing systems*, 20, 2007.

[8] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

[9] Yuansi Chen, Chi Jin, and Bin Yu. Stability and convergence trade-off of iterative optimization algorithms. *arXiv preprint arXiv:1804.01619*, 2018.

[10] 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.

[11] Edgar Dobriban, Stefan Wager, et al. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018.

[12] Andre Elisseeff, Theodoros Evgeniou, Massimiliano Pontil, and Leslie Pack Kaelbing. Stability of randomized learning algorithms. *Journal of Machine Learning Research*, 6(1), 2005.

[13] 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.

[14] Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pages 1832–1841. PMLR, 2018.

[15] Jeff Haochen and Suvrit Sra. Random shuffling beats sgd after finite epochs. In *International Conference on Machine Learning*, pages 2624–2633. PMLR, 2019.

[16] Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pages 1225–1234. PMLR, 2016.

[17] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

[18] 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). *arXiv preprint arXiv:1710.09430*, 2017.

[19] 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.

[20] Ilja Kuzborskij and Christoph Lampert. Data-dependent stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 2815–2824. PMLR, 2018.

[21] Yunwen Lei, Ting Hu, and Ke Tang. Generalization performance of multi-pass stochastic gradient descent with convex loss functions. *J. Mach. Learn. Res.*, 22:25–1, 2021.

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

[23] Siyuan Ma, Raef Bassily, and Mikhail Belkin. The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning. In *International Conference on Machine Learning*, pages 3325–3334. PMLR, 2018.

[24] Eric Moulines and Francis Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in neural information processing systems*, 24, 2011.

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

[26] Deanna Needell, Rachel Ward, and Nati Srebro. Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. *Advances in neural information processing systems*, 27, 2014.

[27] Gergely Neu and Lorenzo Rosasco. Iterate averaging as regularization for stochastic gradient descent. In *Conference On Learning Theory*, pages 3222–3242. PMLR, 2018.

[28] 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.

[29] Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Early stopping and non-parametric regression: an optimal data-dependent stopping rule. *The Journal of Machine Learning Research*, 15(1):335–366, 2014.

[30] Lorenzo Rosasco and Silvia Villa. Learning with incremental iterative regularization. *Advances in Neural Information Processing Systems*, 28, 2015.

[31] Itay Safran and Ohad Shamir. How good is sgd with random shuffling? In *Conference on Learning Theory*, pages 3250–3284. PMLR, 2020.

[32] Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. *Advances in neural information processing systems*, 23, 2010.

[33] Alexander Tsigler and Peter L Bartlett. Benign overfitting in ridge regression. *arXiv preprint arXiv:2009.14286*, 2020.

[34] Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1195–1204. PMLR, 2019.

[35] Sharan Vaswani, Aaron Mishkin, Issam Laradji, Mark Schmidt, Gauthier Gidel, and Simon Lacoste-Julien. Painless stochastic gradient: Interpolation, line-search, and convergence rates. *Advances in neural information processing systems*, 32, 2019.

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

[37] Yuting Wei, Fanny Yang, and Martin J Wainwright. Early stopping for kernel boosting algorithms: A general analysis with localized complexities. *Advances in Neural Information Processing Systems*, 30, 2017.

[38] Denny Wu and Ji Xu. On the optimal weighted \ℓ_2 regularization in overparameterized linear regression. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 10112–10123. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/72e6d3238361fe70f22fb0ac624a7072-Paper.pdf`.

[39] 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. *arXiv preprint arXiv:2110.06198*, 2021.

[40] Ji Xu and Daniel J Hsu. On the number of variables to use in principal component regression. *Advances in neural information processing systems*, 32, 2019.

[41] Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

[42] Yikai Zhang, Wenjia Zhang, Sammy Bald, Vamsi Pingali, Chao Chen, and Mayank Goswami. Stability of sgd: Tightness analysis and improved bounds. *arXiv preprint arXiv:2102.05274*, 2021.

[43] Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, and Sham Kakade. Benign overfitting of constant-stepsize sgd for linear regression. In *Conference on Learning Theory*, pages 4633–4635. PMLR, 2021.

[44] Difan Zou, Jingfeng Wu, Quanquan Gu, Dean P Foster, Sham Kakade, et al. The benefits of implicit regularization from sgd in least squares problems. *Advances in Neural Information Processing Systems*, 34, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] The reported mean behaviors in our experiments are sufficient to support our theoretical results
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] Resources are not a concern for our purpose of numerically verifying our theoretical results
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Proof Sketch for Theorem 4.3

Recall that $\widehat{\mathbf{w}} = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{y} = \mathbf{X}^\top\mathbf{A}^{-1}\mathbf{y}$, where $\mathbf{A} := \mathbf{X}\mathbf{X}^\top$ is the gram matrix. Then we can reformulate $\widehat{\mathbf{w}}_t$ by

$$\widehat{\mathbf{w}}_t = \widehat{\mathbf{w}} - (\mathbf{I} - \eta\mathbf{\Sigma})^t(\widehat{\mathbf{w}}_0 - \widehat{\mathbf{w}}) = \big(\mathbf{I} - (\mathbf{I} - \eta\mathbf{\Sigma})^t\big)\mathbf{X}^\top\mathbf{A}^{-1}\mathbf{y} = \mathbf{X}^\top\big(\mathbf{I} - (\mathbf{I} - \eta n^{-1}\mathbf{A})^t\big)\mathbf{A}^{-1}\mathbf{y}.$$

Denote $\widetilde{\mathbf{A}} := \mathbf{A}\big(\mathbf{I} - (\mathbf{I} - \eta n^{-1}\mathbf{A})^t\big)^{-1}$, the excess risk of $\widehat{\mathbf{w}}_t$ is

$$\mathcal{E}(\widehat{\mathbf{w}}_t) = \frac{1}{2}\big\|\mathbf{X}^\top\widetilde{\mathbf{A}}^{-1}\mathbf{y} - \mathbf{w}^*\big\|_{\mathbf{H}}^2 = \underbrace{\frac{1}{2}\big\|\mathbf{w}^*\big(\mathbf{I} - \mathbf{X}^\top\widetilde{\mathbf{A}}^{-1}\mathbf{X}\big)\big\|_{\mathbf{H}}^2}_{\text{BiasError}} + \underbrace{\frac{1}{2}\big\|\mathbf{X}^\top\widetilde{\mathbf{A}}^{-1}\boldsymbol{\epsilon}\big\|_{\mathbf{H}}^2}_{\text{VarError}}. \tag{A.1}$$

The remaining proof will be relates the excess risk of early stopped GD to that of ridge regression with certain regularization parameters. In particular, note that the excess risk of the ridge regression solution with parameter λ is $\frac{1}{2}\|\mathbf{X}^\top(\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{y} - \mathbf{w}^*\|_{\mathbf{H}}^2$. Then it remains to show the relationship between $\widetilde{\mathbf{A}}$ and $\mathbf{A} + \lambda\mathbf{I}$, which is illustrated in the following lemma.

Lemma A.1 *There is a constant $c > 0$ such that for every $\eta \le c/\lambda_1$ and $t > 0$, it holds that* $\frac{1}{2}\big(\mathbf{A} + \frac{n}{\eta t}\mathbf{I}\big) \preceq \widetilde{\mathbf{A}} \preceq \mathbf{A} + \frac{2n}{t\eta} \cdot \mathbf{I}$.

Then, the lower bound of $\widetilde{\mathbf{A}}$ will be applied to prove the upper bound of variance error of GD, as shown in (A.1), which is at most four times the variance error achieved by the ridge regression with $\lambda = n/(\eta t)$. The upper bound of $\widetilde{\mathbf{A}}$ will be applied to prove the upper bound of the bias error of GD, which is at most the bias error achieved by ridge regression with $\lambda = 2n/(\eta t)$. Finally, we can apply the prior work [33, Theorem 1] on the excess risk analysis for ridge regression to complete the proof for bounding the bias and variance errors separately.

B Risk Bound for the Fluctuation Error

We first state the following properties of the operators $\mathcal{G}$, $\mathcal{M}$, and $\widetilde{\mathcal{M}}$, which are essential in the subsequent analysis:

- **PSD mapping:** for every PSD matrix $\mathbf{J}$, $\mathcal{M} \circ \mathbf{J}$, $(\mathcal{M} - \widetilde{\mathcal{M}}) \circ \mathbf{J}$ and $\mathcal{G} \circ \mathbf{J}$ are all PSD matrices.
- **Commutative property:** for two PSD matrices $\mathbf{B}_1$ and $\mathbf{B}_2$, we have

$$\langle \mathcal{G} \circ \mathbf{B}_1, \mathbf{B}_2\rangle = \langle \mathbf{B}_1, \mathcal{G} \circ \mathbf{B}_2\rangle,\ \langle \mathcal{M} \circ \mathbf{B}_1, \mathbf{B}_2\rangle = \langle \mathbf{B}_1, \mathcal{M} \circ \mathbf{B}_2\rangle,\ \langle \widetilde{\mathcal{M}} \circ \mathbf{B}_1, \mathbf{B}_2\rangle = \langle \mathbf{B}_1, \widetilde{\mathcal{M}} \circ \mathbf{B}_2\rangle$$

B.1 Proof of Inequality (5.4)

Lemma B.1 *The fluctuation error satisfies*

$$\text{FluctuationError} \le \frac{\eta^2}{2} \cdot \sum_{k=0}^{t-1}\langle \mathcal{M} \circ \mathcal{G}^{t-1-k} \circ \mathbf{H}, \mathbf{E}_k\rangle.$$

Proof. [Proof of Lemma B.1] By Lemma 5.3, we have

$$\text{FluctuationError} = \frac{\eta^2}{2} \cdot \sum_{k=0}^{t-1} \langle \mathcal{G}^{t-1-k} \circ (\mathcal{M} - \widetilde{\mathcal{M}}) \circ \mathbf{E}_k, \mathbf{H} \rangle.$$

Then note that $\mathcal{M}$, $\mathcal{M} - \widetilde{\mathcal{M}}$ and $\mathcal{G}$ are the PSD mapping. Then we have

$$\mathcal{G}^{t-1-k} \circ (\mathcal{M} - \widetilde{\mathcal{M}}) \circ \mathbf{E}_k \preceq \mathcal{G}^{t-1-k} \circ \mathcal{M} \circ \mathbf{E}_k$$

for all $k \geq 0$. Further using the commutative property of $\mathcal{G}$ and $\mathcal{M}$, we have

$$\langle \mathcal{G}^{t-1-k} \circ \mathcal{M} \circ \mathbf{E}_k, \mathbf{H} \rangle = \langle \mathcal{M} \circ \mathcal{G}^{t-1-k} \circ \mathbf{H}, \mathbf{E}_k \rangle.$$

This completes the proof.

■

B.2 Proof of Lemma 5.2

We first present the following two useful lemmas.

Lemma B.2 (Theorem 9 in Bartlett et al. [4]) *There is an absolute constant c such that for any $\delta \in (0, 1)$ with probability at least $1 - \delta$,*

$$\|\mathbf{\Sigma} - \mathbf{H}\|_2 \leq c\|\mathbf{H}\|_2 \cdot \max\bigg\{\sqrt{\frac{r(\mathbf{H})}{n}}, \frac{r(\mathbf{H})}{n}, \sqrt{\frac{\log(1/\delta)}{n}}, \frac{\log(1/\delta)}{n}\bigg\},$$

where $r(\mathbf{H}) = \sum_i \lambda_i / \lambda_1$.

Lemma B.3 (Lemma 22 in [4]) *There is a universal constant c such that for any independent, mean zero, σ-subexponential random variables $\xi_1, \ldots, \xi_n$, any $\mathbf{a} = (a_1, \ldots, a_n)$ and any $t \geq 0$,*

$$\mathbb{P}\bigg(\bigg|\sum_{i=1}^n a_i \xi_i\bigg| \geq t\bigg) \leq 2\exp\bigg[-c\min\bigg(\frac{t^2}{\sigma^2\|\mathbf{a}\|_2^2}, \frac{t}{\sigma\|\mathbf{a}\|_\infty}\bigg)\bigg].$$

Proof. [Proof of Lemma 5.2] Note that $(1-x)^k \leq 1/[x(k+1)]$ for all $k > 0$ and $x \in (0, 1)$, we have

$$(\mathbf{I} - \eta\mathbf{\Sigma})^k \mathbf{\Sigma} (\mathbf{I} - \eta\mathbf{\Sigma})^k = \mathbf{\Sigma}(\mathbf{I} - \eta\mathbf{\Sigma})^{2k} \preceq \frac{1}{2(k+1)\eta} \cdot \mathbf{I}.$$

Besides, we also have $\mathbf{\Sigma}(\mathbf{I} - \eta\mathbf{\Sigma})^{2k} \preceq \mathbf{\Sigma}$. This implies that

$$\Theta_1 \leq \min\bigg\{\mathbf{x}_i^\top \mathbf{\Sigma}\mathbf{x}_i, \frac{\|\mathbf{x}_i\|_2^2}{2(k+1)\eta}\bigg\} \leq \min\bigg\{\|\mathbf{\Sigma}\|_2 \cdot \|\mathbf{x}_i\|_2^2, \frac{\|\mathbf{x}_i\|_2^2}{2(k+1)\eta}\bigg\}. \tag{B.1}$$

Then applying Lemma B.2 and using the assumption that $\lambda_1 = \Theta(1)$, we have

$$\|\mathbf{\Sigma}\|_2 \lesssim \|\mathbf{H}\|_2.$$

Besides, by Assumption 3.1, we have

$$\|\mathbf{x}_i\|_2^2 = \sum_i \lambda_i \cdot z_i^2$$

where z_i is independent 1-subgaussian random variable and satisfies $\mathbb{E}[z_i^2] = 1$. Therefore, applying Lemma B.3 we can get with probability $1 - \delta$,

$$\|\mathbf{x}_i\|_2^2 \lesssim \sum_i \lambda_i + \max\bigg\{\log(1/\delta) \cdot \lambda_1, \sqrt{\log(1/\delta)\sum_i \lambda_i^2}\bigg\}.$$

Setting $\delta = 1/\text{poly}(n)$ and applying union bound over all $i \in [n]$, we can get with probability at least $1 - 1/\text{poly}(n)$, it holds that $\|\mathbf{x}_i\|_2^2 \leq \log(n) \cdot \text{tr}(\mathbf{H})$ for all $i \in [n]$. Putting this into (B.1) completes the proof.

■

B.3 Proof of Lemma 5.3

We first provide the following useful facts and lemmas.

Fact B.4 (Part of Lemma 8 in Bartlett et al. [4]) *The gram matrix $\mathbf{A} = \mathbf{X}\mathbf{X}^\top$ can be decomposed by*

$$\mathbf{A} = \sum_i \lambda_i \mathbf{z}_i \mathbf{z}_i^\top,$$

where $\mathbf{z}_i \in \mathbb{R}^n$ are independent 1-subgaussian random vector satisfying $\mathbb{E}[\|\mathbf{z}_i\|_2^2] = n$.

Fact B.5 *Assume $n < d$ and the gram matrix $\mathbf{A}$ is of full-rank, then it holds that*

$$\mathbf{X}(\mathbf{I}_d - \eta\mathbf{\Sigma})^k = (\mathbf{I}_n - \eta n^{-1}\mathbf{A})^k \mathbf{X}.$$

Proof. [Proof of Fact B.5] Note that $\mathbf{X} \in \mathbb{R}^{n\times d}$, consider its SVD decomposition $\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top$, where $\mathbf{U} \in \mathbb{R}^{n\times n}$, $\mathbf{V} \in \mathbb{R}^{d\times d}$ and $\mathbf{\Lambda} \in \mathbb{R}^{n\times d}$. Then we have $\mathbf{\Sigma} = n^{-1}\mathbf{X}^\top\mathbf{X} = n^{-1}\mathbf{V}\mathbf{\Lambda}^\top\mathbf{\Lambda}\mathbf{V}^\top$, which implies that

$$\mathbf{X}(\mathbf{I} - \eta\mathbf{\Sigma})^k = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top\mathbf{V}(\mathbf{I}_d - \eta n^{-1}\mathbf{\Lambda}^\top\mathbf{\Lambda})^k\mathbf{V}^\top = \mathbf{U}\mathbf{\Lambda}(\mathbf{I}_d - \eta n^{-1}\mathbf{\Lambda}^\top\mathbf{\Lambda})^k\mathbf{V}^\top.$$

Additionally, it is easy to verify that $\mathbf{\Lambda}(\mathbf{I}_d - \eta n^{-1}\mathbf{\Lambda}^\top\mathbf{\Lambda}) = (\mathbf{I}_n - \eta n^{-1}\mathbf{\Lambda}\mathbf{\Lambda}^\top)\mathbf{\Lambda}$. Therefore, it follows that

$$\mathbf{X}(\mathbf{I} - \eta\mathbf{\Sigma})^k = \mathbf{U}\mathbf{\Lambda}(\mathbf{I}_d - \eta n^{-1}\mathbf{\Lambda}^\top\mathbf{\Lambda})^k\mathbf{V}^\top = \mathbf{U}(\mathbf{I}_n - \eta n^{-1}\mathbf{\Lambda}\mathbf{\Lambda}^\top)^k\mathbf{\Lambda}\mathbf{V}^\top = (\mathbf{I}_n - \eta\mathbf{A})^k\mathbf{X},$$

where the last equality follows from the fact that $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{\Lambda}^\top\mathbf{U}^\top$. This completes the proof. ■

Lemma B.6 *Let $\mathbf{u} \in \mathcal{S}^{n-1}$ be a uniformly random unit vector, then for any fixed PSD matrix $\mathbf{\Theta} \in \mathbb{R}^{n\times n}$, with probability at least $1 - 1/\mathrm{poly}(n)$, it holds that*

$$\mathbf{u}^\top\mathbf{\Theta}\mathbf{u} \lesssim \frac{\log(n)}{n} \cdot \mathrm{tr}(\mathbf{\Theta}).$$

Proof. We first consider a Gaussian random vector $\mathbf{v} \sim N(0, \mathbf{I}_n/n)$, then it is clear that we can reformulate it as $\mathbf{v} = r \cdot \mathbf{u}$, where $\mathbf{u}$ is a uniformly random unit vector and $\mathbb{E}[r] = 1$. Note that nr^2 follows $\chi^2(n)$ distribution, then by standard concentration result for sub-exponential random variable [36], we have with probability at least $1 - e^{-cn}$ for some small constant $c > 0$ that $r \geq 1/2$. Moreover, let $\mathbf{\Theta} = \sum_i \mu_i \mathbf{z}_i\mathbf{z}_i^\top$ be the eigen-decomposition of $\mathbf{\Theta}$, we have

$$n\mathbf{v}^\top\mathbf{\Theta}\mathbf{v} - \mathrm{tr}(\mathbf{\Theta}) = \sum_{i=1}^n \mu_i[n(\mathbf{z}_i^\top\mathbf{v})^2 - 1] := \sum_{i=1}^n \mu_i\xi_i$$

where $\xi_i \sim \chi^2(1) - 1$ distribution, which is 1-subexponential. Then applying Lemma B.3, we have with probability at least $1 - 2e^{-x}$ such that

$$\sum_{i=1}^n \mu_i\xi_i \leq C \cdot \max\left(x\mu_1, \sqrt{x\sum_{i=1}^n \mu_i^2}\right)$$

holds for some constant C.

Combining the previous results, we have with probability at least $1 - e^{cn} - 2e^{-x}$,

$$\mathbf{u}^\top\mathbf{\Theta}\mathbf{u} = r^{-1}\mathbf{v}^\top\mathbf{\Theta}\mathbf{v} \leq \frac{2}{n}\left[\mathrm{tr}(\mathbf{\Theta}) + C \cdot \max\left(x\mu_1, \sqrt{x\sum_{i=1}^n \mu_i^2}\right)\right].$$

Further note that $\sum_{i=1}^n \mu_i^2, \mu_1 \leq \mathrm{tr}^2(\mathbf{\Theta})$, then setting $x = C'\log(n)$ for some absolute constant C', we have with probability at least $1 - 1/\mathrm{poly}(n)$,

$$\mathbf{u}^\top\mathbf{\Theta}\mathbf{u} = r^{-1}\mathbf{v}^\top\mathbf{\Theta}\mathbf{v} \leq \frac{C''\log(n)}{n} \cdot \mathrm{tr}(\mathbf{\Theta})$$

for some absolute constant C''. This completes the proof.

■

Lemma B.7 *For any $k^* \in [d]$, with probability at least $1 - 1/\mathrm{poly}(n)$, it holds that*

$$\mathrm{tr}(\mathbf{A}(\mathbf{I}_n - \eta n^{-1}\mathbf{A})^{2k}) \lesssim \frac{nk^*}{(k+1)\eta} + n\log(n) \cdot \sum_{i>k^*} \lambda_i.$$

Proof. Let $\mu_1, \ldots, \mu_n$ be the sorted (in descending order) eigenvalues of $\mathbf{A}$, then we have

$$\mathrm{tr}\left(\mathbf{A}(\mathbf{I}_n - \eta n^{-1}\mathbf{A})^{2k})\right) = \sum_{i=1}^{n} \mu_i \cdot (1 - \eta n^{-1}\mu_i)^{2k} \leq \sum_{i=1}^{n} \min\left\{\frac{n}{2(k+1)\eta}, \mu_i\right\}, \tag{B.2}$$

where the inequality follows from the fact that $(1-x)^k \leq 1/[(k+1)x]$ for all $x \in (0,1)$ and $k > 0$. Additionally, by Fact B.4 we have

$$\mathbf{A} = \sum_i \lambda_i \mathbf{z}_i \mathbf{z}_i^\top,$$

where $\{\mathbf{z}_i\}_{i=1,\ldots,n}$ are i.i.d. 1-subgaussian random vectors satisfying $\mathbb{E}[\mathbf{z}_i] = 0$ and $\mathbb{E}[\|\mathbf{z}_i\|_2^2] = n$. Then define

$$\mathbf{A}_k := \sum_{i>k} \lambda_i \mathbf{z}_i \mathbf{z}_i^\top, \tag{B.3}$$

and

$$\mathbf{A}_k = \sum_{i=1}^{n} \mu_i(\mathbf{A}_k)\mathbf{u}_i\mathbf{u}_i^\top$$

be its eigen-decomposition. Then note that $\mathbf{A} - \mathbf{A}_k + \sum_{i=1}^{j} \mu_i(\mathbf{A}_k)\mathbf{u}_i\mathbf{u}_i^\top$ has rank at most $k+j$, thus there must exist a linear space $\mathcal{L}$ of dimension $n-k-j$ (that is orthogonal to $\{\mathbf{z}_i\}_{i=1,\ldots,k}$ and $\{\mathbf{u}_i\}_{i=1}^{j}$) such that for all $\mathbf{v} \in \mathcal{L}$,

$$\mathbf{v}^\top \mathbf{A}\mathbf{v} \leq \mathbf{v}^\top \mu_1\left(\mathbf{A}_k - \sum_{i=1}^{j} \mu_i(\mathbf{A}_k)\mathbf{u}_i\mathbf{u}_i^\top\right)\mathbf{v} = \mathbf{v}^\top \mu_{j+1}(\mathbf{A}_k)\mathbf{v}.$$

This implies that for any $k \in [n]$ and $j \in [n-k]$, it holds that

$$\mu_{k+j}(\mathbf{A}) \leq \mu_j(\mathbf{A}_k),$$

and thus

$$\sum_{i=k+1}^{n} \mu_i \leq \sum_{i=1}^{n+1-i} \mu_i(\mathbf{A}_k) \leq \mathrm{tr}(\mathbf{A}_k). \tag{B.4}$$

Moreover, by the definition of $\mathbf{A}_k$ in (B.3), we have

$$\mathrm{tr}(\mathbf{A}_k) = \sum_{i>k} \lambda_i \|\mathbf{z}_i\|_2^2.$$

Then note that $\|\mathbf{z}_i\|_2^2/n - 1$ is 1-subexponential, by Lemma B.3, we have with probability at least $1 - 2e^{-x}$

$$\mathrm{tr}(\mathbf{A}_k) \leq n\sum_{i>k} \lambda_i + C \cdot n \cdot \max\left(x\lambda_{k+1}, \sqrt{x\sum_{i>k}\lambda_i^2}\right).$$

for some absolute constant C. Then setting $x = \Theta(\log(n))$ and using the fact that $\sum_{i>k} \lambda_i^2 \leq (\sum_{i>k} \lambda_i)^2$, we have with probability at least $1 - 1/\mathrm{poly}(n)$,

$$\mathrm{tr}(\mathbf{A}_k) \lesssim n\log(n) \cdot \sum_{i>k} \lambda_i. \tag{B.5}$$

Putting (B.5) into (B.4) and further applying (B.2), we have for any $k^* \in [n]$, with probability at least $1 - 1/\mathrm{poly}(n)$

$$\mathrm{tr}\left(\mathbf{A}(\mathbf{I}_n - \eta n^{-1}\mathbf{A})^{2k})\right) \leq \sum_{i=1}^{k^*} \frac{n}{2(k+1)\eta} + \mathrm{tr}(\mathbf{A}_k) \lesssim \frac{nk^*}{(k+1)\eta} + n\log(n) \cdot \sum_{i>k^*} \lambda_i.$$

This completes the proof. ■

Proof. [Proof of Lemma 5.3] Recalling the formula of Θ_2, we have

$$\Theta_2 = \mathbf{x}_i^\top (\mathbf{I} - \eta\mathbf{\Sigma})^k (\mathbf{H} - \mathbf{\Sigma})(\mathbf{I} - \eta\mathbf{\Sigma})^k \mathbf{x}_i.$$

Moreover, note that $\mathbf{x}_i$ can be rewritten as $\mathbf{x}_i = \mathbf{e}_i^\top \mathbf{X}$, where $\mathbf{e}_i \in \mathbb{R}^n$ and $\mathbf{X} \in \mathbb{R}^{n\times d}$. Then

$$\begin{aligned}\Theta_2 &= \mathbf{e}_i^\top \mathbf{X}(\mathbf{I} - \eta\mathbf{\Sigma})^k (\mathbf{H} - \mathbf{\Sigma})(\mathbf{I} - \eta\mathbf{\Sigma})^k \mathbf{X}^\top \mathbf{e}_i \\ &\leq \|\mathbf{e}_i^\top \mathbf{X}(\mathbf{I} - \eta\mathbf{\Sigma})^k\|_2^2 \cdot \|\mathbf{H} - \mathbf{\Sigma}\|_2. \end{aligned} \tag{B.6}$$

Then by Fact B.5, we have

$$\begin{aligned}\|\mathbf{e}_i^\top \mathbf{X}(\mathbf{I} - \eta\mathbf{\Sigma})^k\|_2^2 &= \|\mathbf{e}_i^\top (\mathbf{I}_n - \eta n^{-1}\mathbf{A})^k \mathbf{X}\| \\ &= \mathbf{e}_i^\top (\mathbf{I}_n - \eta n^{-1}\mathbf{A})^k \mathbf{X}\mathbf{X}^\top (\mathbf{I}_n - \eta n^{-1}\mathbf{A})^k \mathbf{e}_i \\ &= \mathbf{e}_i^\top \mathbf{A}(\mathbf{I}_n - \eta n^{-1}\mathbf{A})^{2k} \mathbf{e}_i.\end{aligned}$$

Note that $\mathbf{e}_i$ is independent of the randomness of $\mathbf{A}$ and the eigenvectors of $\mathbf{A}$ is rotation invariant. Specifically, note that $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{\Lambda}^\top\mathbf{U}^\top$, where $\mathbf{U} \in \mathbb{R}^{n\times n}$ is an orthonormal matrix and $\mathbf{\Lambda}\mathbf{\Lambda}^\top \in \mathbb{R}^{n\times n}$ is an diagonal matrix. Then we consider the conditional distribution $\mathbb{P}(\mathbf{A}|\mathbf{\Lambda}\mathbf{\Lambda}^\top)$, which can be viewed as a distribution over the orthonormal matrix $\mathbf{U}$, denoted by $\mathbb{P}(\mathbf{U})$. Then note that $\mathbf{U}$ can also be understood as a rotation matrix when operated on an vector, and using Fact B.4, we have for any rotation matrix $\mathbf{P}$, it holds that

$$\mathbf{P}\mathbf{A}\mathbf{P}^\top = \sum_i \lambda_i \mathbf{P}\mathbf{z}_i\mathbf{z}_i^\top\mathbf{P}^\top$$

which has the same distribution of $\mathbf{A} = \sum_i \lambda_i \mathbf{z}_i\mathbf{z}_i^\top$ since $\mathbf{P}\mathbf{z}_i$ and $\mathbf{z}_i$ have the same distribution. Therefore, it can be verified that for any different orthonormal matrices $\mathbf{U}_1$ and $\mathbf{U}_2$ and let $\mathbf{P} = \mathbf{U}_2\mathbf{U}_1^\top$, which is also an orthonormal matrix, we have

$$\mathbb{P}(\mathbf{U}_1\mathbf{\Lambda}\mathbf{\Lambda}^\top\mathbf{U}_1^\top|\mathbf{\Lambda}\mathbf{\Lambda}^\top) = \mathbb{P}(\mathbf{P}\mathbf{U}_1\mathbf{\Lambda}\mathbf{\Lambda}^\top\mathbf{U}_1^\top\mathbf{P}^\top|\mathbf{\Lambda}\mathbf{\Lambda}^\top) = \mathbb{P}(\mathbf{U}_2\mathbf{\Lambda}\mathbf{\Lambda}^\top\mathbf{U}_2^\top|\mathbf{\Lambda}\mathbf{\Lambda}^\top).$$

This implies that $\mathbb{P}(\mathbf{U}_1) = \mathbb{P}(\mathbf{U}_2)$ for any $\mathbf{U}_1 \neq \mathbf{U}_2$. Therefore, we can conclude that $\mathbb{P}(\mathbf{U})$ is an uniform distribution over the entire class of orthonormal matrices. Then note that

$$\mathbf{A}(\mathbf{I}_n - \eta n^{-1}\mathbf{A})^{2k} = \mathbf{P}\big(\mathbf{\Lambda}\mathbf{\Lambda}^\top(\mathbf{I} - n^{-1}\eta\mathbf{\Lambda}\mathbf{\Lambda}^\top)^{2k}\big)\mathbf{P}^\top.$$

Then for any fixed i, using the fact that $\mathbf{P}$ is a uniformly random rotation matrix, we have $\mathbf{P}^\top\mathbf{e}_i$ is a random unit vector in $\mathcal{S}^{n-1}$. Then applying Lemmas B.6 and B.7, and taking union bound over $i \in [n]$, we have with probability at least $1 - 1/\mathrm{poly}(n)$,

$$\begin{aligned}\mathbf{e}_i^\top \mathbf{A}(\mathbf{I}_n - \eta n^{-1}\mathbf{A})^{2k}\mathbf{e}_i &\lesssim \frac{\log(n)}{n} \cdot \mathrm{tr}\left(\mathbf{A}(\mathbf{I}_n - \eta n^{-1}\mathbf{A})^{2k}\right) \\ &\lesssim \log(n) \cdot \left(\frac{k^*}{(k+1)\eta} + \log(n) \cdot \sum_{i>k^*} \lambda_i\right).\end{aligned} \tag{B.7}$$

Finally, applying Lemma B.2 and setting $\delta = 1/\mathrm{poly}(n)$, we have

$$\|\mathbf{H} - \mathbf{\Sigma}\|_2 \lesssim \sqrt{\frac{\log(n)}{n}}. \tag{B.8}$$

Putting (B.8) and (B.7) into (B.6), we can obtain

$$\Theta_2 \leq \|\mathbf{e}_i^\top \mathbf{X}(\mathbf{I} - \eta\mathbf{\Sigma})^k\|_2^2 \cdot \|\mathbf{H} - \mathbf{\Sigma}\|_2 \lesssim \frac{\log^{5/2}(n)}{n^{1/2}} \cdot \left(\frac{k^*}{(k+1)\eta} + \sum_{i>k^*} \lambda_i\right),$$

which completes the proof. ■

B.4 Completing the analysis for fluctuation error: Proof of Theorem 4.2

Combining the established upper bounds on Θ_1 and Θ_2 in Lemmas 5.2 and 5.3 gives the following lemma.

Lemma B.8 *If the stepsize satisfies $\gamma \le 1/(c\,\mathrm{tr}(\mathbf{H}))$ for some absolute constant c, then with probability at least $1 - 1/\mathrm{poly}(n)$, there exists an absolute constant C such that*

$$\mathcal{M} \circ \mathcal{G}^k \circ \mathbf{H} \preceq C \cdot \left[\log(n) \cdot \min\left\{ \frac{1}{(k+1)\eta}, \|\mathbf{H}\|_2 \right\} \cdot \mathrm{tr}(\mathbf{H}) + \frac{\log^{5/2}(n)}{n^{1/2}} \cdot \left(\frac{k^*}{(k+1)\eta} + \sum_{i>k^*} \lambda_i \right) \right] \cdot \mathbf{\Sigma}.$$

Lemma B.9 *For any $t > 0$, if the stepsize satisfies $\eta \le 1/(c\,\mathrm{tr}(\mathbf{H})\log(t))$ for some absolute constant c, then it holds that*

$$\sum_{k=0}^{t-1} \langle \mathbf{\Sigma}, \mathbf{E}_k \rangle \lesssim \frac{1}{\eta} \cdot \langle \mathbf{I} - (\mathbf{I} - \eta\mathbf{\Sigma})^t, \mathbf{E}_0 \rangle,$$

$$\sum_{k=0}^{t-1} \frac{\langle \mathbf{\Sigma}, \mathbf{E}_k \rangle}{t-k} \lesssim \frac{1}{\eta t} \big\langle (\mathbf{I} - (\mathbf{I} - \eta\mathbf{\Sigma})^t), \mathbf{E}_0 \big\rangle + \log(t) \big\langle (\mathbf{I} - \eta\mathbf{\Sigma})^t \mathbf{\Sigma}, \mathbf{E}_0 \big\rangle.$$

Proof. [Proof of Lemma B.9] In this part we seek to bound $\sum_{k=0}^{t-1} \big\langle \mathbf{\Sigma}, \mathbf{E}_k \big\rangle$ and $\sum_{k=0}^{t-1} \frac{\langle \mathbf{\Sigma}, \mathbf{E}_k \rangle}{t-k}$ in separate. By (5.2), we can get

$$\begin{aligned}
\langle \mathbf{\Sigma}, \mathbf{E}_t \rangle &\le \langle \mathbf{\Sigma}, \mathcal{G}^t \circ \mathbf{E}_0 \rangle + \eta^2 \sum_{k=0}^{t-1} \langle \mathbf{\Sigma}, \mathcal{G}^{t-1-k} \circ \mathcal{M} \circ \mathbf{E}_k \rangle \\
&= \langle \mathcal{G}^t \circ \mathbf{\Sigma}, \mathbf{E}_0 \rangle + \eta^2 \sum_{k=0}^{t-1} \langle \mathcal{M} \circ \mathcal{G}^{t-1-k} \circ \mathbf{\Sigma}, \mathbf{E}_k \rangle \\
&= \langle (\mathbf{I} - \eta\mathbf{\Sigma})^{2t} \mathbf{\Sigma}, \mathbf{E}_0 \rangle + \eta^2 \sum_{k=0}^{t-1} \langle \mathcal{M} \circ \big((\mathbf{I} - \eta\mathbf{\Sigma})^{2(t-1-k)} \mathbf{\Sigma} \big), \mathbf{E}_k \rangle. \qquad \text{(B.9)}
\end{aligned}$$

Note that $(\mathbf{I} - \eta\mathbf{\Sigma})^{2(t-1-k)} \mathbf{\Sigma} \preceq \frac{1}{\eta(t-k)} \mathbf{I}$, and $\mathcal{M} \circ \mathbf{I} \preceq c\,\mathrm{tr}(\mathbf{H}) \mathbf{\Sigma}$ for some absolute constant c, we then have the following by (B.9)

$$\langle \mathbf{\Sigma}, \mathbf{E}_t \rangle \le \langle (\mathbf{I} - \eta\mathbf{\Sigma})^{2t} \mathbf{\Sigma}, \mathbf{E}_0 \rangle + c\eta\,\mathrm{tr}(\mathbf{H}) \sum_{k=0}^{t-1} \frac{\langle \mathbf{\Sigma}, \mathbf{E}_k \rangle}{t-k}. \qquad \text{(B.10)}$$

We now bound $\sum_{k=0}^{t-1} \langle \mathbf{\Sigma}, \mathbf{E}_k \rangle$ by recursively applying (B.10) to establish

$$\begin{aligned}
\sum_{k=0}^{t-1} \langle \mathbf{\Sigma}, \mathbf{E}_k \rangle &\le \langle \sum_{k=0}^{t-1} (\mathbf{I} - \eta\mathbf{\Sigma})^{2k} \mathbf{\Sigma}, \mathbf{E}_0 \rangle + c\eta\,\mathrm{tr}(\mathbf{H}) \sum_{k=0}^{t-1} \sum_{i=0}^{k-1} \frac{\langle \mathbf{\Sigma}, \mathbf{E}_i \rangle}{k-i} \\
&\le \frac{1}{\eta} \langle \mathbf{I} - (\mathbf{I} - \eta\mathbf{\Sigma})^t, \mathbf{E}_0 \rangle + 2c\eta\,\mathrm{tr}(\mathbf{H}) \log(t) \sum_{i=0}^{t-1} \langle \mathbf{\Sigma}, \mathbf{E}_i \rangle,
\end{aligned}$$

and conclude that

$$\begin{aligned}
\sum_{k=0}^{t-1} \langle \mathbf{\Sigma}, \mathbf{E}_k \rangle &\le \frac{1}{1 - 2c\eta\,\mathrm{tr}(\mathbf{H})\log(t)} \cdot \frac{1}{\eta} \cdot \langle \mathbf{I} - (\mathbf{I} - \eta\mathbf{\Sigma})^t, \mathbf{E}_0 \rangle \qquad \text{(B.11)} \\
&\le C \cdot \frac{1}{\eta} \cdot \langle \mathbf{I} - (\mathbf{I} - \eta\mathbf{\Sigma})^t, \mathbf{E}_0 \rangle. \qquad \text{(B.12)}
\end{aligned}$$

Similarly, we then bound $\sum_{k=0}^{t-1} \frac{\langle \mathbf{\Sigma}, \mathbf{E}_k \rangle}{t-k}$ by recursively applying (B.10) to establish

$$\sum_{k=0}^{t-1} \frac{\langle \mathbf{\Sigma}, \mathbf{E}_k \rangle}{t-k} \le \langle \sum_{k=0}^{t-1} \frac{(\mathbf{I} - \eta\mathbf{\Sigma})^{2k} \mathbf{\Sigma}}{t-k}, \mathbf{E}_0 \rangle + c\eta\,\mathrm{tr}(\mathbf{H}) \sum_{k=0}^{t-1} \sum_{i=0}^{k-1} \frac{\langle \mathbf{\Sigma}, \mathbf{E}_i \rangle}{(t-k)(k-i)}$$

$$\leq \langle \sum_{k=0}^{t-1} \frac{(\mathbf{I}-\eta\mathbf{\Sigma})^{2k}\mathbf{\Sigma}}{t-k}, \mathbf{E}_0\rangle + 2c\eta\,\mathrm{tr}(\mathbf{H})\log(t)\sum_{i=0}^{t-1}\frac{\langle \mathbf{\Sigma}, \mathbf{E}_i\rangle}{t-i},$$

so we can conclude that

$$\sum_{k=0}^{t-1}\frac{\langle \mathbf{\Sigma}, \mathbf{E}_k\rangle}{t-k} \leq \frac{1}{1-2c\eta\,\mathrm{tr}(\mathbf{H})\log(t)}\langle \sum_{k=0}^{t-1}\frac{(\mathbf{I}-\eta\mathbf{\Sigma})^{2k}\mathbf{\Sigma}}{t-k}, \mathbf{E}_0\rangle \tag{B.13}$$

$$\lesssim \sum_{k=0}^{t-1}\frac{(\mathbf{I}-\eta\mathbf{\Sigma})^{2k}\mathbf{\Sigma}}{t-k}, \mathbf{E}_0\rangle \tag{B.14}$$

$$\lesssim \left(\frac{1}{\eta t}\big\langle (\mathbf{I}-(\mathbf{I}-\eta\mathbf{\Sigma})^t), \mathbf{E}_0\big\rangle + \log(t)\big\langle (\mathbf{I}-\eta\mathbf{\Sigma})^t\mathbf{\Sigma}, \mathbf{E}_0\big\rangle\right), \tag{B.15}$$

where the last inequality is due to

$$\sum_{k=0}^{t-1}\frac{(\mathbf{I}-\eta\mathbf{\Sigma})^{2k}\mathbf{\Sigma}}{t-k} \lesssim \frac{1}{\eta t}(\mathbf{I}-(\mathbf{I}-\eta\mathbf{\Sigma})^t) + \log(t)\cdot(\mathbf{I}-\eta\mathbf{\Sigma})^t\mathbf{\Sigma}.$$

■

Lemma B.10 *For any $t \geq 0$ and $\eta \leq 1/(c\,\mathrm{tr}(\mathbf{H})\log(t))$ for some absolute constant c, it holds that*

$$\big\langle \mathbf{I}-(\mathbf{I}-\eta\mathbf{\Sigma})^t, \mathbf{E}_0\big\rangle \leq \min\big\{\|\widehat{\mathbf{w}}\|_2^2, t\eta\cdot\langle \mathbf{\Sigma}, \mathbf{E}_0\rangle\big\}$$
$$t\eta\cdot\big\langle (\mathbf{I}-\eta\mathbf{\Sigma})^t\mathbf{\Sigma}, \mathbf{E}_0\big\rangle \leq \min\big\{\|\widehat{\mathbf{w}}\|_2^2, t\eta\cdot\langle \mathbf{\Sigma}, \mathbf{E}_0\rangle\big\}$$

Proof. According to the definition of $\mathbf{E}_t$ and applying zero initialization $\mathbf{w}_0 = \mathbf{0}$, then we have $\mathbf{E}_0 = \widehat{\mathbf{w}}\widehat{\mathbf{w}}^\top \preceq \|\widehat{\mathbf{w}}\|_2^2\cdot\mathbf{I}$. Moreover, note that our choice of stepsize guarantees that $\mathbf{I}-\eta\mathbf{\Sigma}$ is a PSD matrix, we have

$$\mathbf{I}-(\mathbf{I}-\eta\mathbf{\Sigma})^t \preceq \mathbf{I}, \quad \mathbf{I}-(\mathbf{I}-\eta\mathbf{\Sigma})^t \preceq t\eta\mathbf{\Sigma}, \quad (\mathbf{I}-\eta\mathbf{\Sigma})^t\mathbf{\Sigma} \preceq \mathbf{\Sigma}, \quad (\mathbf{I}-\eta\mathbf{\Sigma})^t\mathbf{\Sigma} \preceq \frac{1}{t\eta}\cdot\mathbf{I}.$$

Then it follows that

$$\big\langle \mathbf{I}-(\mathbf{I}-\eta\mathbf{\Sigma})^t, \mathbf{E}_0\big\rangle \leq \min\big\{\langle \mathbf{I}, \mathbf{E}_0\rangle, t\eta\cdot\langle \mathbf{\Sigma}, \mathbf{E}_0\rangle\big\} = \min\big\{\|\widehat{\mathbf{w}}\|_2^2, t\eta\cdot\langle \mathbf{\Sigma}, \mathbf{E}_0\rangle\big\}$$
$$\big\langle (\mathbf{I}-\eta\mathbf{\Sigma})^t\mathbf{\Sigma}, \mathbf{E}_0\big\rangle \leq \min\left\{\langle \mathbf{\Sigma}, \mathbf{E}_0\rangle, \frac{1}{t\eta}\cdot\langle \mathbf{I}, \mathbf{E}_0\rangle\right\} = \min\left\{\langle \mathbf{\Sigma}, \mathbf{E}_0\rangle, \frac{\|\widehat{\mathbf{w}}\|_2^2}{t\eta}\right\}.$$

This completes the proof.

■

Now we are ready to complete the proof of Theorem 4.2. **Proof.** [Proof of Theorem 4.2] By Lemma B.1, we have

$$\underbrace{\mathrm{FluctuationError}}_{*} \leq \frac{\eta^2}{2}\cdot\sum_{k=0}^{t-1}\langle \mathcal{M}\circ\mathcal{G}^{t-1-k}\circ\mathbf{H}, \mathbf{E}_k\rangle.$$

Additionally, by Lemma B.8, we further have

$$(*) \lesssim \eta^2\cdot\sum_{k=0}^{t-1}\left[\frac{\log(n)}{(t-k)\eta}\cdot\mathrm{tr}(\mathbf{H}) + \frac{\log^{5/2}(n)}{n^{1/2}}\cdot\left(\frac{k^*}{(t-k)\eta} + \sum_{i>k^*}\lambda_i\right)\right]\cdot\langle \mathbf{\Sigma}, \mathbf{E}_k\rangle$$
$$\lesssim \eta\cdot\left(\log(n)\,\mathrm{tr}(\mathbf{H}) + \frac{k^*\log^{5/2}(n)}{n^{1/2}}\right)\cdot\sum_{k=0}^{t-1}\frac{\langle \mathbf{\Sigma}, \mathbf{E}_k\rangle}{t-k} + \eta^2\cdot\frac{\log^{5/2}(n)}{n^{1/2}}\cdot\sum_{i>k^*}\lambda_i\cdot\sum_{k=0}^{t-1}\langle \mathbf{\Sigma}, \mathbf{E}_k\rangle.$$

Then applying Lemma B.9, we can further obtain

$$(*) \lesssim \eta\cdot\left(\log(n)\,\mathrm{tr}(\mathbf{H}) + \frac{k^*\log^{5/2}(n)}{n^{1/2}}\right)\cdot\left(\frac{1}{\eta t}\big\langle (\mathbf{I}-(\mathbf{I}-\eta\mathbf{\Sigma})^t), \mathbf{E}_0\big\rangle + \log(t)\big\langle (\mathbf{I}-\eta\mathbf{\Sigma})^t\mathbf{\Sigma}, \mathbf{E}_0\big\rangle\right)$$

$$
\begin{aligned}
&+ \eta \cdot \frac{\log^{5/2}(n)}{n^{1/2}} \cdot \sum_{i>k^*} \lambda_i \cdot \langle \mathbf{I} - (\mathbf{I} - \eta\mathbf{\Sigma})^t, \mathbf{E}_0 \rangle \\
= &\left(\frac{\log(n)\operatorname{tr}(\mathbf{H})}{t} + \frac{\log^{5/2}(n)}{n^{1/2}t} \cdot (k^* + \eta t \sum_{i>k^*} \lambda_i) \right) \cdot \big\langle (\mathbf{I} - (\mathbf{I} - \eta\mathbf{\Sigma})^t), \mathbf{E}_0 \big\rangle \\
&+ \eta \log(t) \cdot \left(\log(n)\operatorname{tr}(\mathbf{H}) + \frac{k^* \log^{5/2}(n)}{n^{1/2}} \right) \cdot \big\langle (\mathbf{I} - \eta\mathbf{\Sigma})^t \mathbf{\Sigma}, \mathbf{E}_0 \big\rangle \\
\lesssim &\left[\log(t) \cdot \left(\frac{\operatorname{tr}(\mathbf{H})\log(n)}{t} + \frac{k^* \log^{5/2}(n)}{n^{1/2}t} \right) + \frac{\log^{5/2}(n)\eta}{n^{1/2}} \cdot \sum_{i>k^*} \lambda_i \right) \right] \cdot \min\big\{ \|\widehat{\mathbf{w}}\|_2^2, t\eta \cdot \langle \mathbf{\Sigma}, \mathbf{E}_0 \rangle \big\}.
\end{aligned}
$$

where the last inequality follows from Lemma B.10.

■

C Risk bounds for Gradient Descent with Early Stopping

C.1 Proof of Lemma A.1

Proof. [Proof of Lemma A.1] For the first inequality, note that

$$
\mathbf{I} - (\mathbf{I} - \eta n^{-1}\mathbf{A})^t \preceq \begin{cases} \mathbf{I}; \\ n^{-1}\eta t \mathbf{A}, \end{cases}
$$

we then obtain

$$
\widetilde{\mathbf{A}} := \mathbf{A}\big(\mathbf{I} - (\mathbf{I} - \eta n^{-1}\mathbf{A})^t\big)^{-1} \succeq \begin{cases} \mathbf{A}; \\ \frac{n}{\eta t}\mathbf{I}. \end{cases}
$$

Therefore

$$
\widetilde{\mathbf{A}} \succeq \frac{1}{2}\big(\mathbf{A} + \frac{n}{\eta t}\mathbf{I}\big).
$$

For the second inequality, note that

$$
\widetilde{\mathbf{A}} - \mathbf{A} = \mathbf{A}(\mathbf{I} - \eta n^{-1}\mathbf{A})^t \big[\mathbf{I} - (\mathbf{I} - \eta n^{-1}\mathbf{A})^t\big]^{-1}.
$$

Then it suffices to consider the scalar function $f(x) := nx(1-\eta x)^t / \big[1 - (1-\eta x)^t\big]$. Then we consider two cases: (1) $t\eta x \geq \log(2)$ and (2) $t\eta x < \log(2)$. For the first case, it is clear that

$$
\frac{nx(1-\eta x)^t}{1-(1-\eta x)^t} \leq \frac{n \cdot 1/(t\eta)}{1 - 1/2} = \frac{2n}{t\eta},
$$

where we use the inequality $(1-\eta x)^t x \leq 1/(t\eta)$ in the first inequality. For the case of $t\eta x < \log(2)$, we have $(1-\eta x)^t \leq 1 - \eta x t/2$ and thus

$$
\frac{nx(1-\eta x)^t}{1-(1-\eta x)^t} \leq \frac{nx}{\eta x t/2} = \frac{2n}{t\eta}.
$$

Combining the about results in two cases, we have $f(x) \leq 2n/(t\eta)$ and thus

$$
\widetilde{\mathbf{A}} = \mathbf{A} + \mathbf{A}(\mathbf{I} - \eta n^{-1}\mathbf{A})^t \big[\mathbf{I} - (\mathbf{I} - \eta n^{-1}\mathbf{A})^t\big]^{-1} \preceq \mathbf{A} + \frac{2n}{t\eta} \cdot \mathbf{I}.
$$

This completes the proof of the second inequality.

■

Then, the lower bound of $\widetilde{\mathbf{A}}$ will be applied to prove the upper bound of variance error of GD, as shown in (A.1), which is at most four times the variance error achieved by the ridge regression with $\lambda = n/(\eta t)$. The upper bound of $\widetilde{\mathbf{A}}$ will be applied to prove the upper bound of the bias error of GD, which is at most the bias error achieved by ridge regression with $\lambda = 2n/(\eta t)$. Finally, we can apply the prior work [33, Theorem 1] on the excess risk analysis for ridge regression to complete the proof for bounding the bias and variance errors separately. The detailed proofs are provided as follows.

C.2 Variance Error

Lemma C.1 *For any stepsize $\gamma \le c/\operatorname{tr}(\mathbf{H})$ for some absolute constant c and any $k^* \in [d]$, with probability at least $1 - 1/\mathrm{poly}(n)$,*

$$\mathbb{E}_{\boldsymbol{\epsilon}}[\mathrm{VarError}] \lesssim \frac{k^*}{n} + \frac{n}{\big(n/(\eta t) + \sum_{i>k^*} \lambda_i\big)^2} \cdot \sum_{i>k^*} \lambda_i^2$$

Proof. By (A.1), we have

$$\mathbb{E}_{\boldsymbol{\epsilon}}[\mathrm{VarError}] := \big\|\mathbf{X}^\top \widetilde{\mathbf{A}}^{-1}\boldsymbol{\epsilon}\big\|_{\mathbf{H}}^2 \lesssim \operatorname{tr}\big(\mathbf{X}\mathbf{H}\mathbf{X}^\top \widetilde{\mathbf{A}}^{-2}\big) \lesssim \operatorname{tr}\Big(\mathbf{X}\mathbf{H}\mathbf{X}^\top \Big(\mathbf{A} + \frac{n}{\eta t}\mathbf{I}\Big)^{-2}\Big), \tag{C.1}$$

where the last inequality is by Lemma A.1. One finds that (C.1) corresponds to the variance error of ridge regression in [33] for $\lambda = \frac{n}{\eta t}$. Then by Theorem 1 in Tsigler and Bartlett [33], one immediately obtains a bound for GD variance error:

$$\mathbb{E}_{\boldsymbol{\epsilon}}[\mathrm{VarError}] \lesssim \frac{k^*}{n} + \frac{n}{\big(n/(\eta t) + \sum_{i>k^*} \lambda_i\big)^2} \cdot \sum_{i>k^*} \lambda_i^2,$$

where

$$k^* := \min\Big\{k : n\lambda_{k+1} \le \frac{n}{\eta t} + \sum_{i>k} \lambda_i\Big\}.$$

Setting $\widetilde{\lambda} = n/(\eta t) + \sum_{i>k^*} \lambda_i$ completes the proof. ∎

C.3 Bias Error

Lemma C.2 *Assume the ground truth $\mathbf{w}^*$ follows a Gaussian Prior $\mathbf{w}^* \sim \mathcal{N}(0, \omega^2 \cdot \mathbf{I})$. Then for any stepsize $\gamma \le c/\operatorname{tr}(\mathbf{H})$ for some absolute constant c and any $k^* \in [d]$, with probability at least $1 - 1/\mathrm{poly}(n)$,*

$$\mathbb{E}_{\mathbf{w}^*}[\mathrm{BiasError}] \lesssim \omega^2 \cdot \Big(\frac{\widetilde{\lambda}^2}{n^2} \cdot \sum_{i \le k^*} \frac{1}{\lambda_i} + \sum_{i>k^*} \lambda_i\Big).$$

Proof. Note that given the ground truth $\mathbf{w}^*$, the bias error is

$$\mathrm{BiasError} := \|\mathbf{H}^{\frac{1}{2}}\big(\mathbf{I} - \mathbf{X}^\top \widetilde{\mathbf{A}}^{-1}\mathbf{X}\big)\mathbf{w}^*\|_2^2.$$

Further note that

$$\mathbf{w}^* \sim \mathcal{N}(0, \omega^2 \cdot \mathbf{I}_d),$$

then taking expectation over $\mathbf{w}^*$ gives

$$\begin{aligned}
\mathbb{E}_{\mathbf{w}^*}[\mathrm{BiasError}] &= \mathbb{E}_{\mathbf{w}^*}\big[\|\mathbf{H}^{\frac{1}{2}}\big(\mathbf{I} - \mathbf{X}^\top \widetilde{\mathbf{A}}^{-1}\mathbf{X}\big)\mathbf{w}^*\|_2^2\big] \\
&= \omega^2 \cdot \operatorname{tr}\big(\mathbf{H}\big(\mathbf{I} - \mathbf{X}^\top \widetilde{\mathbf{A}}^{-1}\mathbf{X}\big)^2\big) \\
&\le \omega^2 \cdot \underbrace{\operatorname{tr}\Big(\mathbf{H}\Big(\mathbf{I} - \mathbf{X}^\top \Big(\mathbf{A} + \frac{2n}{t\eta}\Big)^{-2}\mathbf{X}\Big)^2\Big)}_{*}
\end{aligned}$$

where the last inequality is by Lemma A.1 and that $\mathbf{A}$ commutes with $\widetilde{\mathbf{A}}$. Moreover, note that the quantity $(*)$ is actually the expected bias error of the ridge regression solution with the regularization parameter $2n/(t\eta)$. Therefore, by Theorem 1 in Tsigler and Bartlett [33], we have

$$\begin{aligned}
(*) &\lesssim \mathbb{E}_{\mathbf{w}^* \sim \mathcal{N}(\mathbf{0}, \mathbf{I})}\Bigg[\Bigg(\frac{2n/(\eta t) + \sum_{i>k^*} \lambda_i}{n}\Bigg)^2 \cdot \|\mathbf{w}^*_{0:k^*}\|^2_{\mathbf{H}^{-1}_{0:k^*}} + \|\mathbf{w}^*_{k^*:\infty}\|^2_{\mathbf{H}_{k^*:\infty}}\Bigg] \\
&\eqsim \frac{\widetilde{\lambda}^2}{n^2} \cdot \sum_{i \le k^*} \frac{1}{\lambda_i} + \sum_{i>k^*} \lambda_i,
\end{aligned}$$

where

$$k^* := \min\Big\{k : n\lambda_{k+1} \le \frac{n}{\eta t} + \sum_{i>k} \lambda_i\Big\},$$

and $\widetilde{\lambda} = n/(\eta t) + \sum_{i>k^*} \lambda_i$. This completes the proof. ∎

C.4 Proof of Theorem 4.3

Proof. [Proof of Theorem 4.3] The proof can be completed by combining Lemmas C.1 and C.2. ■

D Proof of Corollaries

D.1 Proof of Corollary 4.4

The following lemma will be useful in the proof.

Lemma D.1 *Assume* $\mathbf{w}^* \sim \mathcal{N}(\mathbf{0}, \omega^2 \cdot \mathbf{I})$ *and* $\mathbf{w}_0 = \mathbf{0}$*, then*

$$\mathbb{E}_{\mathbf{w}^*,\boldsymbol{\epsilon}}[\langle \mathbf{E}_0, \boldsymbol{\Sigma}\rangle] \lesssim \omega^2 \cdot \log(n) \cdot \mathrm{tr}(\mathbf{H}) + \sigma^2.$$
$$\mathbb{E}_{\mathbf{w}^*,\boldsymbol{\epsilon}}[\|\widehat{\mathbf{w}}\|_2^2] = n\omega^2 + \sigma^2\,\mathrm{tr}(\mathbf{A}^{-1}).$$

Proof. Applying the formula of $\widehat{\mathbf{w}}$ and the initialization $\mathbf{w}_0 = \mathbf{0}$, we have

$$\langle \mathbf{E}_0, \boldsymbol{\Sigma}\rangle = \langle \mathbf{X}^\top \mathbf{A}^{-1}\mathbf{y}(\mathbf{X}^\top \mathbf{A}^{-1}\mathbf{y})^\top, \boldsymbol{\Sigma}\rangle = \frac{1}{n}\|\mathbf{y}\|_2^2 = \frac{1}{n}\|\mathbf{X}\mathbf{w}^* + \boldsymbol{\epsilon}\|_2^2 \le \frac{2}{n}\|\mathbf{X}\mathbf{w}^*\|_2^2 + \frac{2}{n}\|\boldsymbol{\epsilon}\|_2^2,$$

where the last inequality follows from Young's inequality. Note that $\boldsymbol{\epsilon}$ is a combination of n independent random variables with variance σ^2, we have $\mathbb{E}[\|\boldsymbol{\epsilon}\|_2^2] = n\sigma^2$. Besides, regarding the first term, we have with probability at least $1 - 1/\mathrm{poly}(n)$,

$$\mathbb{E}[\|\mathbf{X}\mathbf{w}^*\|_2^2] = \omega^2 \cdot \mathrm{tr}(\mathbf{X}\mathbf{X}^\top) \lesssim \omega^2 \cdot n \cdot \mathrm{tr}(\mathbf{H}).$$

Combining the above results immediately gives

$$\mathbb{E}_{\mathbf{w}^*,\boldsymbol{\epsilon}}[\langle \mathbf{E}_0, \boldsymbol{\Sigma}\rangle] \lesssim \omega^2 \cdot \mathrm{tr}(\mathbf{H}) + \sigma^2.$$

Moreover, note that $\widehat{\mathbf{w}} = \mathbf{X}^\top \mathbf{A}^{-1}\mathbf{y} = \mathbf{X}^\top \mathbf{A}^{-1}(\mathbf{X}\mathbf{w}^* + \boldsymbol{\epsilon})$, we have

$$\begin{aligned}
\mathbb{E}_{\mathbf{w}^*,\boldsymbol{\epsilon}}[\|\widehat{\mathbf{w}}\|_2^2] &= \mathbb{E}_{\mathbf{w}^*,\boldsymbol{\epsilon}}\big[\,\mathrm{tr}\left(\mathbf{X}^\top \mathbf{A}^{-1}(\mathbf{X}\mathbf{w}^* + \boldsymbol{\epsilon})(\mathbf{X}\mathbf{w}^* + \boldsymbol{\epsilon})^\top \mathbf{A}^{-1}\mathbf{X}\right)\big] \\
&= \mathbb{E}_{\mathbf{w}^*,\boldsymbol{\epsilon}}\big[\,\mathrm{tr}\left(\mathbf{X}^\top \mathbf{A}^{-1}\mathbf{X}\mathbf{w}^*\mathbf{w}^{*\top}\mathbf{X}^\top \mathbf{A}^{-1}\mathbf{X}\right) + \mathrm{tr}\left(\mathbf{X}^\top \mathbf{A}^{-1}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top \mathbf{A}^{-1}\mathbf{X}\right)\big] \\
&= \omega^2\,\mathrm{tr}\left(\mathbf{X}^\top \mathbf{A}^{-1}\mathbf{X}\mathbf{X}^\top \mathbf{A}^{-1}\mathbf{X}\right) + \sigma^2\,\mathrm{tr}(\mathbf{X}^\top \mathbf{A}^{-2}\mathbf{X}),
\end{aligned}$$

where the last equality is due to $\mathbf{w}^* \sim \mathcal{N}(\mathbf{0}, \omega^2\mathbf{I})$ and $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Then note that $\mathbf{A} = \mathbf{X}\mathbf{X}^\top$, we have

$$\mathbb{E}_{\mathbf{w}^*,\boldsymbol{\epsilon}}[\|\widehat{\mathbf{w}}\|_2^2] = \omega^2\,\mathrm{tr}(\mathbf{I}_n) + \sigma^2\,\mathrm{tr}(\mathbf{A}^{-1}) = n\omega^2 + \sigma^2\,\mathrm{tr}(\mathbf{A}^{-1}).$$

This completes the proof.

■

Proof. [Proof of Corollary 4.4] Plugging Lemma D.1 into Theorem 4.2 and then combining Theorems 4.2 and 4.3 completes the proof.

■

D.2 Proof of Corollary 4.5

Proof. [Proof of Corollary 4.5] First, note that $k^\dagger$ can be arbitrarily chosen, we will first pick $k^\dagger = t\eta/\log(t\eta)^\beta$, which leads to $\sum_{i>k^\dagger} = \log(t\eta)^{1-\beta}$. Then Corollary 4.4 implies that

$$\begin{aligned}
&\mathbb{E}_{\mathrm{SGD},\mathbf{w}^*,\boldsymbol{\epsilon}}\big[\mathrm{FlutuationError}(\mathbf{w}_t)\big] \\
&\lesssim \frac{\eta}{\log(1/\eta)} \cdot \left[\log(t)\log(n) + \frac{\log^{5/2}(n)t\eta}{n^{1/2}} \cdot \log(t\eta)^{1-\beta}\right] \cdot \min\left\{\frac{n\omega^2 + \sigma^2\,\mathrm{tr}(\mathbf{A}^{-1})}{t\eta}, \omega^2\,\mathrm{tr}(\mathbf{H}) + \sigma^2\right\}.
\end{aligned}$$

It is clear that when $t \to \infty$, we have

$$\mathbb{E}_{\mathrm{SGD},\mathbf{w}^*,\boldsymbol{\epsilon}}\big[\mathrm{FlutuationError}(\mathbf{w}_t)\big] \lesssim \frac{\eta(n\omega^2 + \sigma^2\,\mathrm{tr}(\mathbf{A}^{-1}))}{\log(1/\eta)} \cdot \left[\frac{\log(t)}{t\eta}\log(n) + \frac{\log^{5/2}(n)}{n^{1/2}} \cdot \log(t\eta)^{1-\beta}\right]$$

$$\to 0. \tag{D.1}$$

Then we will move on to the GD error in Corollary 4.4, we can get that

$$\lambda_{k^*} \eqsim \frac{1}{\eta t} + \frac{1}{n}\sum_{i>k^*}\lambda_i.$$

Plugging the fact that $\lambda_i = i^{-1}\log(i+1)^{-\beta}$, the above equality implies that

$$(k^*)^{-1}\log(k^*)^{-\beta} \eqsim \frac{1}{\eta t} + \frac{1}{n}\log(k^*)^{1-\beta},$$

which further leads to

$$k^* = \min\left\{\frac{t\eta}{\log(t\eta)^\beta}, \frac{n}{\log(n)}\right\}.$$

Note that when characterizing the upper bound of SGD, we can pick k^* arbitrarily. Therefore, we will consider two cases accordingly: (1) $t\eta/\log(t\eta)^\beta \le n/\log(n)$; and (2) $t\eta/\log(t\eta)^\beta > n/\log(n)$. For the first case, we will pick $k^* = t\eta/\log(t\eta)^\beta$ and get that

$$\widetilde{\lambda} \eqsim \frac{n}{t\eta} + \log(t\eta)^{1-\beta}.$$

Then given the value of k^*, we can further obtain that

$$\begin{aligned}
&\sum_{i\le k^*}\frac{1}{\lambda_i} = \sum_{i\le k^*} i\log(i+1)^\beta \eqsim (k^*)^2\log(k^*)^\beta \eqsim \frac{(t\eta)^2}{\log(t\eta)^\beta}\\
&\sum_{i\le k^*}\lambda_i \eqsim \log(t\eta)^{1-\beta}\\
&\sum_{i>k^*}\lambda_i^2 = \sum_{i>k^*}\frac{1}{i^2\log(i+1)^{2\beta}} \eqsim \frac{1}{k^*\log(k^*)^{2\beta}} \eqsim \frac{1}{t\eta\log(t\eta)^\beta}.
\end{aligned}$$

Therefore, we can get that

$$\frac{\widetilde{\lambda}^2}{n^2}\cdot\sum_{i\le k^*}\frac{1}{\lambda_i} + \sum_{i>k^*}\lambda_i \eqsim \log(t\eta)^{1-\beta}, \quad \frac{k^*}{n} + \frac{n}{\widetilde{\lambda}^2}\sum_{i>k^*}\lambda_i^2 \lesssim \frac{t\eta}{n\log(t\eta)^\beta}.$$

Then we can get the following according to Corollary 4.4,

$$\begin{aligned}
&\mathbb{E}_{\mathrm{SGD},\mathbf{w}^*,\boldsymbol{\epsilon}}\big[\mathcal{E}(\mathbf{w}_t)\big] \lesssim \omega^2\cdot\log(t\eta)^{1-\beta} + \sigma^2\cdot\frac{t\eta}{n\log(t\eta)^\beta}\\
&+ \frac{\eta}{\log(1/\eta)}\cdot\left[\log(t)\log(n) + \frac{\log^{5/2}(n)t\eta}{n^{1/2}}\cdot\log(t\eta)^{1-\beta}\right]\cdot\min\left\{\frac{n\omega^2+\sigma^2\operatorname{tr}(\mathbf{A}^{-1})}{t\eta}, \omega^2\operatorname{tr}(\mathbf{H})+\sigma^2\right\},
\end{aligned}$$

Taking $t\eta = n$ and set

$$\eta \lesssim \log^{-3}(n)\cdot n^{-1/2},$$

we can immediately get that

$$\mathbb{E}_{\mathrm{SGD},\mathbf{w}^*,\boldsymbol{\epsilon}}[\mathcal{E}(\mathbf{w}_t)] \lesssim \omega^2\cdot\log(n)^{1-\beta} + \sigma^2\log(n)^{-\beta}.$$

For the second case of $t\eta/\log(t\eta)^\beta > n/\log(n)$, we will pick $k^* = n/\log(n)$, which leads to $\widetilde{\lambda} = \log(n)^{1-\beta}$. Then we can get

$$\sum_{i\le k^*}\frac{1}{\lambda_i} \eqsim \frac{n^2}{\log(n)^{2-\beta}}, \quad \sum_{i\le k^*}\lambda_i \eqsim \log(n)^{1-\beta}, \quad \sum_{i>k^*}\lambda_i^2 \eqsim \frac{1}{n\log(n)^{2\beta-1}}.$$

This further leads to

$$\frac{\widetilde{\lambda}^2}{n^2}\cdot\sum_{i\le k^*}\frac{1}{\lambda_i} + \sum_{i>k^*}\lambda_i \eqsim \log(n)^{1-\beta}, \quad \frac{k^*}{n} + \frac{n}{\widetilde{\lambda}^2}\sum_{i>k^*}\lambda_i^2 \eqsim \frac{1}{\log(n)}.$$

Then we can get the following upper bound on the excess risk of SGD:

$$\mathbb{E}_{\mathrm{SGD},\mathbf{w}^*,\boldsymbol{\epsilon}}\big[\mathcal{E}(\mathbf{w}_t)\big] \lesssim \omega^2 \cdot \log(n)^{1-\beta} + \sigma^2 \cdot \frac{1}{\log(n)}$$
$$+ \frac{\eta}{\log(1/\eta)} \cdot \left[\log(t)\log(n) + \frac{\log^{5/2}(n) t\eta}{n^{1/2}} \cdot \log(t\eta)^{1-\beta}\right] \cdot \min\left\{\frac{n\omega^2 + \sigma^2 \operatorname{tr}(\mathbf{A}^{-1})}{t\eta}, \omega^2 \operatorname{tr}(\mathbf{H}) + \sigma^2\right\}.$$

Taking $t \to \infty$ and applying (D.1) give

$$\lim_{t\to\infty} \mathbb{E}_{\mathrm{SGD},\mathbf{w}^*,\boldsymbol{\epsilon}}[\mathcal{E}(\mathbf{w}_t)] \lesssim \omega^2 \cdot \log(n)^{1-\beta} + \sigma^2 \log(n)^{-1}.$$

This completes the proof.

■

D.3 Proof of Corollary 4.6

Proof. [Proof of Corollary 4.6] We will first calculate k^* defined in Corollary 4.4. Note that

$$k^* = \min\left\{k : n\lambda_{k+1} \le \frac{n}{\eta t} + \sum_{i>k} \lambda_i\right\},$$

and $\sum_{i>k} \lambda_i = \sum_{i>k} i^{-1-r} \eqsim k^{-r}$. Then, it can be shown that

$$k^* = (t\eta)^{1/(r+1)}. \tag{D.2}$$

Recall that Corollary 4.4 shows

$$\begin{aligned}
&\mathbb{E}_{\mathrm{SGD},\mathbf{w}^*}[\mathrm{Risk}(\mathbf{w}_t)] \\
&\lesssim \omega^2 \cdot \underbrace{\left(\frac{\widetilde{\lambda}^2}{n^2} \cdot \sum_{i\le k^*} \frac{1}{\lambda_i} + \sum_{i>k^*} \lambda_i\right)}_{I_1} + \sigma^2 \cdot \underbrace{\left(\frac{k^*}{n} + \frac{n}{\widetilde{\lambda}^2} \sum_{i>k^*} \lambda_i^2\right)}_{I_2} + (\omega^2 \operatorname{tr}(\mathbf{H}) + \sigma^2))\eta \\
&\quad \cdot \underbrace{\left[\log(t) \cdot \left(\operatorname{tr}(\mathbf{H})\log(n) + \frac{k^* \log^{5/2}(n)}{n^{1/2}}\right) + \frac{\log^{5/2}(n) t\eta}{n^{1/2}} \cdot \sum_{i>k^*} \lambda_i\right)\right]}_{I_3}.
\end{aligned}$$

Then, applying (D.2) gives

$$\sum_{i>k^*} \lambda_i \eqsim (k^*)^{-r} \eqsim (t\eta)^{-r/(r+1)}, \ \sum_{i>k^*} \lambda_i^2 \eqsim (k^*)^{-2r-1} \eqsim (t\eta)^{-(2r+1)/(r+1)},$$
$$\sum_{i\le k^*} \frac{1}{\lambda_i} \eqsim (k^*)^{r+2} = (t\eta)^{(r+2)/(r+1)}, \ \widetilde{\lambda} \eqsim \frac{n}{t\eta}, \ \operatorname{tr}(\mathbf{H}) \eqsim 1$$

Putting the above into the formula of I_1, I_2, and I_3, we can get

$$\begin{aligned}
I_1 &\lesssim \frac{n^2/(t\eta)^2}{n^2} \cdot (t\eta)^{(r+2)/(r+1)} + (t\eta)^{-r/(r+1)} \eqsim (t\eta)^{-r/(r+1)}; \\
I_2 &\lesssim \frac{(t\eta)^{1/(r+1)}}{n} + \frac{n}{n^2/(t\eta)^2} \cdot (t\eta)^{-(2r+1)/(r+1)} \eqsim \frac{(t\eta)^{1/(r+1)}}{n}; \\
I_3 &\lesssim \log(t) \cdot \left(\log(n) + \frac{(t\eta)^{1/(r+1)} \log^{5/2}(n)}{n^{1/2}}\right) + \frac{\log^{5/2}(n) t\eta}{n^{1/2}} \cdot (t\eta)^{-r/(r+1)}\Big) \\
&\lesssim \log(t) \cdot \left[\log(n) + \frac{\log^{5/2}(n)}{n^{1/2}} \cdot (t\eta)^{1/(r+Corollary1)}\right].
\end{aligned}$$

Combining the above results leads to

$$\mathbb{E}_{\mathrm{SGD},\mathbf{w}^*}[\mathrm{Risk}(\mathbf{w}_t)] \lesssim \omega^2 \cdot (t\eta)^{-r/(r+1)} + \sigma^2 \cdot \frac{(t\eta)^{1/(r+1)}}{n}$$

$$+ (\omega^2 + \sigma^2) \cdot \eta \cdot \log(t) \cdot \left[\log(n) + \frac{\log^{5/2}(n)}{n^{1/2}} \cdot (t\eta)^{1/(r+1)} \right].$$

This completes the proof.

■