Last Iterate Risk Bounds of SGD with Decaying Stepsize for Overparameterized Linear Regression

Jingfeng Wu [*1] **Difan Zou** [*2] **Vladimir Braverman** [1] **Quanquan Gu** [2] **Sham M. Kakade** [3]

Abstract

Stochastic gradient descent (SGD) has been shown to generalize well in many deep learning applications. In practice, one often runs SGD with a *geometrically decaying stepsize*, i.e., a constant initial stepsize followed by multiple geometric stepsize decay, and uses the *last iterate* as the output. This kind of SGD is known to be nearly minimax optimal for classical finite-dimensional linear regression problems (Ge et al., 2019). However, a sharp analysis for the last iterate of SGD in the overparameterized setting is still open. In this paper, we provide a *problem-dependent* analysis on the last iterate risk bounds of SGD with decaying stepsize, for (overparameterized) linear regression problems. In particular, for last iterate SGD with (tail) geometrically decaying stepsize, we prove nearly matching upper and lower bounds on the excess risk. Moreover, we provide an excess risk lower bound for last iterate SGD with polynomially decaying stepsize and demonstrate the advantage of geometrically decaying stepsize in an *instance-wise* manner, which complements the minimax rate comparison made in prior works.

1. Introduction

It is widely observed in practice that modern neural networks trained by *stochastic gradient descent* (SGD) often generalize well (Zhang et al., 2021). In all the successful applications, two ingredients are crucial: (1) an overparameterized model, where the number of parameter excesses the number of training examples (Belkin et al., 2020); and (2) SGD with the *last iterate* as output and with a *decaying stepsize*, e.g., an initially large stepsize, followed by geometrically decaying stepsizes after every certain number of iterations (He et al., 2015). Theoretically, however, it remains largely open to understand the generalization of the *last iterate of SGD* (with a *decaying stepsize*) for learning overparamerized models, even for the arguably simplest setting such as *overparamerized linear regression*.

For linear regression in the classical regime, Ge et al. (2019) showed that last iterate SGD (with geometrically decaying stepsize) can achieve nearly minimax optimal excess risk up to logarithmic factors. However, the results by Ge et al. (2019) cannot be carried over to the overparameterized setting since their excess risk bounds are *dimension-dependent*, which become vacuous when the problem dimension excesses the sample size. There is a fundamental barrier to extend the statistical minimax rate to the overparameterized setting, as the minimax result concerns the *worst instance* in the problem class, while apparently SGD cannot generalize for certain overparameterized linear regression problem (e.g., when the data distribution has an identity covariance and the model parameter is uniformly distributed).

For SGD with iterate averaging, a recent work by Zou et al. (2021b) proved a tight problem-dependent excess risk bound for overparameterized linear regression, which can diminish in the overparameterized setting, provided a sufficiently fast decaying spectrum of the data covaraiance matrix. While Zou et al. (2021b) sharply characterized the generalization of SGD with iterate averaging in the overparameterized setting, their analysis is tailored to the averaged iterate of SGD and is not directly applicable to the last iterate of SGD.

In this paper, in order to explain its success in learning overparameterized models, we provide a tight analysis for the last iterate of SGD that adapts to both of the least-square problem instance and the algorithm configuration.

Contributions. Our first main result is a sharp problem-dependent excess risk bound for the *last iterate SGD with tail geometrically decaying stepsize* (see (3), also Algorithm 1) for linear regression. The derived bound does not depend on the ambient dimension and, instead, depends on the spec-

[*]Equal contribution [1]Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, USA [2]Department of Computer Science, University of California, Los Angeles, CA 90095, USA [3]Department of Computer Science and Department of Statistics, Harvard University, Cambridge, MA 02138, USA. Correspondence to: Vladimir Braverman <vova@cs.jhu.edu>, Quanquan Gu <qgu@cs.ucla.edu>, Sham M. Kakade <sham@seas.harvard.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

Algorithm 1 LAST ITERATE SGD WITH TAIL GEOMETRIC DECAYING STEPSIZE

Require: Initial weight $\mathbf{w}$, initial stepsize γ, total sample size N, first phase length s, decaying phase length K

for $t = 1, \dots, N$ **do**
  if $t > s$ and $(t-s) \bmod K = 0$ **then**
    $\gamma \leftarrow \gamma/2$
  end if
  $\mathbf{w} \leftarrow \mathbf{w} + \gamma(y - \langle \mathbf{w}, \mathbf{x} \rangle)\mathbf{x}$, with a fresh data $(\mathbf{x}, y)$
end for
return $\mathbf{w}$

trum of the data covariance matrix. In particular, the excess risk bound vanishes as long as the data covariance matrix has a fast-decay eigenspectrum, despite of a large ambient dimension in the overparameterized setting. Furthermore, an excess risk lower bound is proved, which shows the upper bound is tight up to absolute constant in terms of variance error, and is nearly tight in terms of bias error. This result recovers the existing minimax bound in the classical regime (Polyak & Juditsky, 1992; Bach & Moulines, 2013) ignoring logarithmic factors, and is comparable to the bounds for SGD with iterate averaging in the overparameterized setting (Zou et al., 2021b).

Our second main result is a comparison between SGD with (1) tail geometrically stepsize-decaying scheme and (2) tail polynomially stepsize-decaying scheme, in an *instance-wise* manner. Our result shows that the variance error of SGD with tail polynomially decaying stepsize is *instance-wise no better* than that of SGD with tail geometrically decaying stepsize, given the same optimization trajectory length (i.e., summation of stepsizes). In contrast, the comparison between these two stepsize schemes made in Ge et al. (2019) only concerns the worst-case result: the worst-case excess risk bound achieved by geometrically decaying stepsize is strictly better than the worst case bound achieved by polynomially decaying stepsize. Thus, their analysis does not rule out the possibility that for some problem instances, polynomially decaying stepsize can generalize better than the geometrically decaying one.

Our analysis follows and extends the operator method for analyzing SGD in linear regression (Dieuleveut et al., 2017; Jain et al., 2017a;b; Neu & Rosasco, 2018; Ge et al., 2019; Zou et al., 2021b). Specifically, we develop a novel, multi-phase analysis for the bias error of the last SGD iterate, which sharpens existing results (see Section 5). We believe our proof technique is of broader interest and can be applied to analyze other variants of SGD such as SGD with momentum.

Notation. We reserve lower-case letters for scalars, lower-case boldface letters for vectors, upper-case boldface letters for matrices, and upper-case calligraphic letters for linear operators on symmetric matrices. For two positive-value functions $f(x)$ and $g(x)$ we write $f(x) \lesssim g(x)$ or $f(x) \gtrsim g(x)$ if $f(x) \leq cg(x)$ or $f(x) \geq cg(x)$ for some absolute constant $c > 0$ respectively. For two vectors $\mathbf{u}$ and $\mathbf{v}$ in a Hilbert space, their inner product is denoted by $\langle \mathbf{u}, \mathbf{v} \rangle$ or equivalently, $\mathbf{u}^\top \mathbf{v}$. For a matrix $\mathbf{A}$, its spectral norm is denoted by $\|\mathbf{A}\|_2$. For two matrices $\mathbf{A}$ and $\mathbf{B}$ of appropriate dimension, their inner product is defined as $\langle \mathbf{A}, \mathbf{B} \rangle := \mathrm{tr}(\mathbf{A}^\top \mathbf{B})$. For a positive semi-definite (PSD) matrix $\mathbf{A}$ and a vector $\mathbf{v}$ of appropriate dimension, we write $\|\mathbf{v}\|_{\mathbf{A}}^2 := \mathbf{v}^\top \mathbf{A} \mathbf{v}$. The Kronecker/tensor product is denoted by $\otimes$. Finally, $\log(\cdot)$ refers to logarithm base 2.

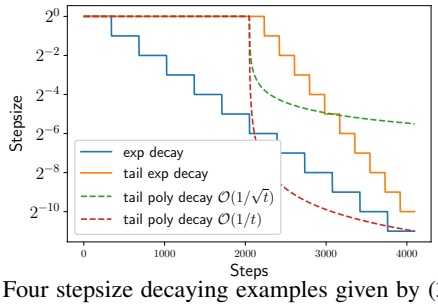


Figure 1. Four stepsize decaying examples given by (3) and (4). $\gamma_0 = 1$ and $N = 4096$. EXP DECAY: (3) with $s = 0$ and $K = \lceil N/\log N \rceil$; TAIL EXP DECAY: (3) with $s = N/2$ and $K = \lceil (N-s)/\log(N-s) \rceil$; TAIL POLY DECAY $\mathcal{O}(1/\sqrt{t})$: (4) with $s = N/2$ and $a = 0.5$; TAIL POLY DECAY $\mathcal{O}(1/t)$: (4) with $s = N/2$ and $a = 1$.

2. Related Work

Problem-Dependent Bounds for Linear Regression. We first discuss a set of dimension-free and problem-dependent bounds for linear regression that are similar to what we show in this paper. Bartlett et al. (2020) proved risk bounds of ordinary least square (OLS) for overparameterized linear regression in terms of the full eigenspectrum of the data covariance matrix, and showed that benign overfitting can occur even when OLS memorizes the training data. Tsigler & Bartlett (2020) extended the benign overfitting result of OLS to ridge regression, and proved diminishing risk bounds for a larger class of least square problems. Zou et al. (2021b) proved problem-dependent risk bounds for constant-stepsize SGD with iterate averaging (and tail-averaging) and compared the algorithmic regularization afforded by SGD with OLS and ridge regression. Our excess risk upper bound (Theorem 4.1) for last iterate SGD is comparable to theirs for SGD with averaging (Theorems 2.1 and 5.1 in Zou et al. (2021b)). Due to this similarity, the benefits of SGD with tail-averaging over ridge regression, as discussed in Zou et al. (2021a), naturally extends to the, more practical, last iterate SGD studied in this paper.

Nonparamatric Bounds for SGD. We then discuss other SGD risk bounds for infinite-dimensional/nonparamatric linear regression (Dieuleveut & Bach, 2015; Lin & Rosasco,

2017; Mücke et al., 2019; Berthier et al., 2020; Varre et al., 2021). Dieuleveut & Bach (2015) only discussed linear regression with data covariance whose spectrum decays polynomially, in contrast our results apply to general data covariance. The works by Berthier et al. (2020); Varre et al. (2021) only dealt with bias error (i.e., they assume no additive label noise), but we provide both variance and bias error bounds. Compared to Lin & Rosasco (2017); Mücke et al. (2019); Berthier et al. (2020); Varre et al. (2021), our results rely on a different set of assumptions: they assume a stronger condition on the optimal model parameter ($\mathbf{w}^*$), which requires $\|\mathbf{H}^{-\alpha}\mathbf{w}^*\|_2$ to be finite for some constant $\alpha > 0$ where $\mathbf{H}$ is the data covariance; though we do not require this, our assumption on the fourth moment operator is stronger (see Assumption 3.2).

Last Iterate SGD with Decaying Stepsize in the Classical Regime. In the finite-dimensional setting, there is a rich literature considering the last iterate SGD with decaying stepsize. For example, polynomially decaying stepsizes are studied in (Dekel et al., 2012; Rakhlin et al., 2011; Lacoste-Julien et al., 2012; Bubeck, 2014), and geometrically decaying stepsizes are considered in (Davis et al., 2019; Ghadimi & Lan, 2012; Hazan & Kale, 2014; Aybat et al., 2019; Kulunchakov & Mairal, 2019; Ge et al., 2019); besides, a recent work by Pan et al. (2021) explored eigenvalue-dependent stepsizes. However, the bounds derived in the aforementioned papers are all *dimension-dependent* and therefore cannot be applied to the overparameterized setting. In this regard, our work can be viewed as a dimension-free, problem-dependent extension of Ge et al. (2019)'s results that are limited to finite-dimensional, worst-case scenarios.

Finally, we would like to refer the reader to Table 1 in Section 4.1 for a detailed comparison between our results and several existing ones (Ge et al., 2019; Bach & Moulines, 2013; Zou et al., 2021b).

3. Problem Setup and Preliminaries

We now formally set up the problem.

High-Dimensional Linear Regression. Let $\mathbf{x}$ be a feature vector in a Hilbert space that can be d-dimensional or countably infinite dimensional, and $y \in \mathbb{R}$ be the response. *Linear regression* concerns the following objective:

$$\min_{\mathbf{w}} L(\mathbf{w}), \text{ where } L(\mathbf{w}) := \frac{1}{2}\mathbb{E}\left(y - \langle \mathbf{w}, \mathbf{x}\rangle\right)^2, \quad (1)$$

where $\mathbf{w}$ is a weight vector to be learned, and the expectation is over an unknown joint distribution $\mathcal{D}$ of $(\mathbf{x}, y)$[1].

SGD. We consider solving (1) using *stochastic gradient descent* (SGD). The weight vector is initialized at $\mathbf{w}_0$ in the Hilbert space; then at the t-th iteration, a fresh data $(\mathbf{x}_t, y_t)$ is drawn independently from the distribution, and the weight vector is updated according to

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \gamma_t(y_t - \langle \mathbf{w}_{t-1}, \mathbf{x}_t\rangle)\mathbf{x}_t,\ t = 1, 2, \ldots, N, \quad (2)$$

where $\gamma_t > 0$ is the stepsize at step t. Our main focus in this paper is *last iterate SGD with decaying stepsize*, which uses a sequence of properly decaying stepsize $(\gamma_t)_{t=1}^N$, and outputs the last iterate $\mathbf{w}_N$. For example, one can use *tail geometrically decaying stepsize* (see also Algorithm 1):

$$\gamma_t = \begin{cases} \gamma_0, & 0 \le t \le s; \\ \gamma_0/2^\ell, & s < t \le N, \ell = \lfloor (t-s)/K \rfloor, \end{cases} \quad (3)$$

where the stepsize is kept as a constant in the first s steps, and is then divided by a factor of 2 every K steps. Figure 1 shows two examples of such stepsize decay schemes. We note that (3) captures the widely used stepsize decaying scheduler in deep learning (He et al., 2015): the stepsize is epoch-wise a constant, and decays geometrically after every certain number of epochs.

Another widely studied variant of SGD is *constant-stepsize* SGD with *averaging*. More specifically, it updates the iterate according to (2) with a constant stepsize, i.e., $\gamma_t = \gamma$, and its final output is an averaging of all iterates ($\frac{1}{N}\sum_{t=0}^{N-1}\mathbf{w}_t$) or only the tail iterates ($\frac{1}{N-s}\sum_{t=s}^{N-1}\mathbf{w}_t$). Compared with last iterate SGD, SGD with averaging is less practical but more theoretically favorable. For a few examples, the risk bounds of SGD with averaging have been studied in both the classical underparameterized regime (Bach & Moulines, 2013; Dieuleveut et al., 2017; Jain et al., 2017a;b; Neu & Rosasco, 2018) and the overparameterized setting (Dieuleveut & Bach, 2015; Zou et al., 2021b).

Next we review a set of assumptions for our analysis.

Assumption 3.1 (Regularity conditions)**.** Denote $\mathbf{H} := \mathbb{E}[\mathbf{x}\mathbf{x}^\top]$, and assume that $\mathbf{H}$ is (entry-wisely) finite and $\mathrm{tr}(\mathbf{H})$ is finite. For convenience, we further assume that $\mathbf{H}$ is strictly positive definite and that $L(\mathbf{w})$ admits a unique global optimum, denoted by $\mathbf{w}^* := \arg\min_{\mathbf{w}} L(\mathbf{w})$.

The condition $\mathbf{H} \succ 0$ is only made for simple presentation; if $\mathbf{H}$ has zero eigenvalues, one can choose $\mathbf{w}^* = \arg\min\{\|\mathbf{w}\|_2 : \mathbf{w} \in \arg\min L(\mathbf{w})\}$, and our results still hold. This argument also holds in a reproducing kernel Hilbert space (Schölkopf et al., 2002).

Assumption 3.2 (Fourth moment conditions)**.** Assume that the fourth moment of $\mathbf{x}$ is finite and:

A There is a constant $\alpha > 0$, such that for every PSD matrix $\mathbf{A}$, we have

$$\mathbb{E}[\mathbf{x}\mathbf{x}^\top\mathbf{A}\mathbf{x}\mathbf{x}^\top] \preceq \alpha \cdot \mathrm{tr}(\mathbf{H}\mathbf{A}) \cdot \mathbf{H}.$$

Clearly, it must hold that $\alpha \ge 1$.

[1] Unless otherwise noted, all expectations in this paper are taken with respect to the joint distribution of $(\mathbf{x}, y)$.

B There is a constant $\beta > 0$, such that for every PSD matrix $\mathbf{A}$, we have

$$\mathbb{E}[\mathbf{x}\mathbf{x}^\top \mathbf{A} \mathbf{x}\mathbf{x}^\top] - \mathbf{H}\mathbf{A}\mathbf{H} \succeq \beta \cdot \mathrm{tr}(\mathbf{H}\mathbf{A}) \cdot \mathbf{H}.$$

To give an example, if $\mathbf{H}^{-\frac{1}{2}}\mathbf{x}$ satisfies Gaussian distribution, then Assumption 3.2 holds with $\alpha = 3$ and $\beta = 1$. More generally, Assumption 3.2A holds for data distributions with a *bounded kurtosis along every direction* (Dieuleveut et al., 2017), i.e., there is a constant $\kappa > 0$ such that

$$\text{for every } \mathbf{v},\ \mathbb{E}[\langle \mathbf{v}, \mathbf{x}\rangle^4] \le \kappa \langle \mathbf{v}, \mathbf{H}\mathbf{v}\rangle^2. \qquad (\text{3.2A'})$$

One can verify that condition (3.2A') (hence Assumption 3.2A) is weaker than assuming a sub-Gaussian tail for the distribution of $\mathbf{H}^{-\frac{1}{2}}\mathbf{x}$ (see Lemma A.1 in Zou et al. (2021b)), where the latter condition is typically made in regression analysis (Hsu et al., 2014; Bartlett et al., 2020; Tsigler & Bartlett, 2020). On the other hand, Assumption 3.2A is stronger than the condition $\mathbb{E}[\mathbf{x}\mathbf{x}^\top \mathbf{x}\mathbf{x}^\top] \preceq R^2\mathbf{H}$ for some constant $R^2 > 0$, as often assumed in many SGD analysis (Bach & Moulines, 2013; Dieuleveut et al., 2017; Jain et al., 2017a;b; Neu & Rosasco, 2018; Ge et al., 2019). We refer the reader to Appendix A for more examples for Assumption 3.2.

Assumption 3.3 (Noise condition)**.** Denote

$$\mathbf{\Sigma} := \mathbb{E}[(y - \langle \mathbf{w}^*, \mathbf{x}\rangle)^2 \mathbf{x}\mathbf{x}^\top], \quad \sigma^2 := \left\|\mathbf{H}^{-\frac{1}{2}}\mathbf{\Sigma}\mathbf{H}^{-\frac{1}{2}}\right\|_2,$$

and assume $\mathbf{\Sigma}$ and σ^2 are finite.

Here $\mathbf{\Sigma}$ is the covariance matrix of the gradient noise at the optimal $\mathbf{w}^*$, and σ^2 characterizes the noise level in that $\mathbf{\Sigma} \preceq \sigma^2\mathbf{H}$. Assumption 3.3 allows the additive noise to be mis-specified (Dieuleveut et al., 2017; Jain et al., 2017b); and in particular, Assumption 3.3 is directly implied by the following Assumption 3.3' for a well-specified linear regression model.

Assumption 3.3' (Well-specified noise)**.** *Assume the response is generated by*

$$y = \langle \mathbf{w}^*, \mathbf{x}\rangle + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2),$$

where ϵ is independent with $\mathbf{x}$.

Additional Notation. Denote the eigen decomposition of the Hessian by $\mathbf{H} = \sum_i \lambda_i \mathbf{v}_i \mathbf{v}_i^\top$, where $(\lambda_i)_{i\ge 1}$ are eigenvalues in a non-increasing order and $(\mathbf{v}_i)_{i\ge 1}$ are the corresponding eigenvectors. We denote $\mathbf{H}_{k^*:k^\dagger} := \sum_{k^*<i\le k^\dagger} \lambda_i \mathbf{v}_i \mathbf{v}_i^\top$, where $0 \le k^* \le k^\dagger$ are two integers, and we allow $k^\dagger = \infty$. For example,

$$\mathbf{H}_{0:k} = \sum_{1\le i\le k} \lambda_i \mathbf{v}_i \mathbf{v}_i^\top, \quad \mathbf{H}_{k:\infty} = \sum_{i>k} \lambda_i \mathbf{v}_i \mathbf{v}_i^\top.$$

Similarly, we denote $\mathbf{I}_{k^*:k^\dagger} := \sum_{k^*<i\le k^\dagger} \mathbf{v}_i \mathbf{v}_i^\top$.

4. Main Results

In this section, we present our main results.

4.1. An Upper Bound

We begin with an excess risk upper bound for last iterate SGD with tail geometrically decaying stepsize.

Theorem 4.1 (An upper bound)**.** *Consider last iterate SGD with stepsize scheme* (3)*. Suppose Assumptions 3.1, 3.2A and 3.3 hold. Let $K := \lceil (N-s)/\log(N-s)\rceil$. Suppose $\gamma_0 < 1/(3\alpha\,\mathrm{tr}(\mathbf{H})\log(s+K))$. Then we have*

$$\mathbb{E}[L(\mathbf{w}_N) - L(\mathbf{w}^*)] \le \mathtt{BiasError} + \mathtt{VarianceError},$$

where

$$\mathtt{BiasError} \lesssim \frac{\left\|(\mathbf{I}-\gamma_0\mathbf{H})^{s+K}(\mathbf{w}_0-\mathbf{w}^*)\right\|^2_{\mathbf{I}_{0:k^*}}}{\gamma_0 K} + \left\|(\mathbf{I}-\gamma_0\mathbf{H})^{s+K}(\mathbf{w}_0-\mathbf{w}^*)\right\|^2_{\mathbf{H}_{k^*:\infty}} + \log(s+K)\cdot\left(\frac{\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{I}_{0:k^\dagger}}}{\gamma_0(s+K)} + \|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{H}_{k^\dagger:\infty}}\right)\cdot\frac{d_{\mathtt{eff}}}{K},$$

and

$$\mathtt{VarianceError} \le \frac{8\sigma^2}{1-\alpha\gamma_0\,\mathrm{tr}(\mathbf{H})}\cdot\frac{d_{\mathtt{eff}}}{K}.$$

Here k^*, $k^\dagger$ are arbitrary indexes, and the effective dimension *is defined by*

$$d_{\mathtt{eff}} := k^* + \gamma_0 K \sum_{k^*<i\le k^\dagger} \lambda_i + \gamma_0^2 K(s+K)\sum_{i>k^\dagger}\lambda_i^2.$$

Moreover, the bound is minimized for $k^* := \max\{k : \lambda_k \ge 1/(\gamma_0 K)\}$ and $k^\dagger := \max\{k : \lambda_k \ge 1/(\gamma_0(s+K))\}$.

The excess risk bound in Theorem 4.1 consists a *bias error* term that stems from the incorrect initialization $\mathbf{w}_0 - \mathbf{w}^* \ne 0$, and a *variance error* term that stems from the presence of additive noise $y - \langle \mathbf{w}^*, \mathbf{x}\rangle \ne 0$. Our bound is *dimension-free* and *problem-dependent*: instead of depending on the ambient dimension d, it depends on the *effective dimension* $d_{\mathtt{eff}}$, which is jointly determined by the problem and the algorithm. In particular, when the eigenspectrum of the data covariance decays fast, the effective dimension $d_{\mathtt{eff}}$ could be much smaller than the ambient dimension d (and sample size N) to enable generalization in the overparameterized scenarios.

For example, let us consider $s = N/2$ and $K = N/(2\log(N/2))$, which corresponds to SGD that starts decaying stepsize after seeing half of the samples. Then the excess risk bound vanishes provided that $d_{\mathtt{eff}} = o(K)$, or in other words,

$$k^* = o\Big(\frac{N}{\log(N)}\Big),\ \sum_{k^*<i\le k^\dagger}\lambda_i = o(1),\ \sum_{i>k^\dagger}\lambda_i^2 = o\Big(\frac{1}{N}\Big).$$

Table 1. A comparison between our result and several existing results. See Section 4.1 for more details.

	Bach & Moulines (2013)	Ge et al. (2019)	Zou et al. (2021b)	Ours
output	averaged iterate	last iterate	averaged iterate	last iterate
initial stepsize	$\gamma \lesssim 1$	$\gamma \lesssim 1$	$\gamma \lesssim 1$	$\gamma \lesssim 1/\log(N)$
effective number of steps ($N_{\mathtt{eff}}$)	N	$\frac{N}{\log(N)}$	N	$\frac{N}{\log(N)}$
effective dimension ($d_{\mathtt{eff}}$)	d	d	$k^* + \gamma^2 N^2 \sum_{i>k^*} \lambda_i^2$	$k^* + \gamma^2 {N_{\mathtt{eff}}}^2 \sum_{i>k^*} \lambda_i^2$
effective noise ($\sigma^2_{\mathtt{eff}}$)	σ^2	σ^2	$\sigma^2 + \frac{\|\mathbf{w}^*\|^2_{\mathbf{I}_{0:k^*}}}{\gamma N_{\mathtt{eff}}} + \|\mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}}$	$\sigma^2 + \log(N)\cdot \left(\frac{\|\mathbf{w}^*\|^2_{\mathbf{I}_{0:k^*}}}{\gamma N_{\mathtt{eff}}} + \|\mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}}\right)$
effective bias error ($\mathtt{Bias}_{\mathtt{eff}}$)	$\frac{\|\mathbf{w}^*\|_2^2}{\gamma N_{\mathtt{eff}}}$	$\frac{d\|\mathbf{w}^*\|_2^2}{\gamma N_{\mathtt{eff}}}$	$\frac{\|\mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k^*}}}{\gamma^2 {N_{\mathtt{eff}}}^2} + \|\mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}}$	$\frac{\|(\mathbf{I}-\gamma\mathbf{H})^{N_{\mathtt{eff}}}\mathbf{w}^*\|^2_{\mathbf{I}_{0:k^*}}}{\gamma N_{\mathtt{eff}}} + \|(\mathbf{I}-\gamma\mathbf{H})^{N_{\mathtt{eff}}}\mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}}$
unified risk bound	$\mathtt{Bias}_{\mathtt{eff}} + \sigma^2_{\mathtt{eff}} \cdot \frac{d_{\mathtt{eff}}}{N_{\mathtt{eff}}}$			

Theorem 4.1 allows the last iterate of SGD to generalize even in the overparameterized regime ($d > N$). Several concrete examples are presented in Corollary 1.

Corollary 1 (Example data distributions)**.** *Under the same conditions as Theorem 4.1, suppose that $s = N/2$, $K = N/(2\log(N/2))$, $\gamma_0 = 1/(4\alpha\,\mathrm{tr}(\mathbf{H})\log(N))$, and $\|\mathbf{w}_0 - \mathbf{w}^*\|_2$ is finite. Recall the eigenspectrum of $\mathbf{H}$ is $(\lambda_k)_{k\geq 1}$.*

1. *If $\lambda_k = k^{-(1+r)}$ for some constant $r > 0$, then the excess risk is $\mathcal{O}\big(N^{\frac{-r}{1+r}} \cdot \log^{\frac{r-1}{1+r}}(N)\big)$.*
2. *If $\lambda_k = k^{-1}\log^{-r}(k+1)$ for some constant $r > 1$, then the excess risk is $\mathcal{O}\big(\log^{-r}(N)\big)$.*
3. *If $\lambda_k = 2^{-k}$, then the excess risk is $\mathcal{O}\big(N^{-1}\log^2(N)\big)$.*

These examples are from Corollary 2.3 in Zou et al. (2021b) for SGD with iterate-averaging (one can verify that their Corollary 2.3 also holds for constant-stepsize SGD with tail-averaging with $s = N/2$). Comparing our Corollary 1 with Corollary 2.3 in Zou et al. (2021b), we can see that the excess risk bounds of last iterate SGD is inferior to that of SGD with averaging by at most polylogarithmic factors.

Reduction to the Classical Regime. It is worthy noting that Theorem 4.1 nearly recovers the minimax optimal bounds (Polyak & Juditsky, 1992; Bach & Moulines, 2013) in the classical regime when $d = o(N)$. In particular, let us set $k^* = k^\dagger = d$, $s = 0$ and $K = N/\log(N)$, then Theorem 4.1 implies:

$$\mathtt{BiasError} \lesssim \frac{\|\mathbf{w}_0 - \mathbf{w}^*\|_2^2 \log(N)}{\gamma_0 N}\Big(1 + \frac{d\log^2(N)}{N}\Big),$$
$$\mathtt{VarianceError} \lesssim \frac{\sigma^2 d \log(N)}{N}.$$

Now choose $\gamma_0 = 1/(4\alpha\,\mathrm{tr}(\mathbf{H})\log\log(N))$ and recall $d = o(N)$, then both the bias and variance errors match the statistical minimax rates (Polyak & Juditsky, 1992; Bach & Moulines, 2013) up to some logarithmic factors.

Comparison with Existing Bounds. Table 1 presents a detailed comparison between our result and several existing results, including Ge et al. (2019) for last iterate SGD and Bach & Moulines (2013); Zou et al. (2021b) for SGD with iterate averaging. To unify notations, we use γ, N, d, $\mathbf{w}^*$, σ^2 to denote the (initial) stepsize, the total number of steps, the ambient dimension, the optimal model parameter and the noise level, respectively. We also use *effective number of steps* as the number of equivalently steps when using constant stepsize (or can be understood as the total optimization length). The *effective dimension* can be understood as the number of useful dimensions (discovered by the algorithm) that contribute to the problem. We also assume all algorithms are initialized from zero ($\mathbf{w}_0 = 0$), without lose of

generality. To be consistent with the algorithmic setting of Ge et al. (2019), we restrict our result to geometric decaying stepsize scheduler ($s = 0$), which decreases the effective number of steps in our result. Table 1 shows that our result generalizes that in Ge et al. (2019) for last iterate SGD to high dimensional setting, and is comparable to that in Zou et al. (2021b) for SGD with iterate averaging ignoring some logarithmic factors.

4.2. A Lower Bound

We complement the above upper bound with a lower bound.

Theorem 4.2 (A lower bound)**.** *Consider last iterate SGD with stepsize scheme* (3)*. Suppose Assumptions 3.1, 3.2B and 3.3' hold. Let $K = (N-s)/\log(N-s)$. Suppose $K \geq 10$ and $\gamma_0 < 1/\lambda_1$. Then we have*

$$\mathbb{E}[L(\mathbf{w}_N) - L(\mathbf{w}^*)] = \frac{1}{2}\texttt{BiasError} + \frac{1}{2}\texttt{VarianceError},$$

where

$$\texttt{BiasError} \geq \left\|(\mathbf{I}-\gamma_0\mathbf{H})^{s+2K}(\mathbf{w}_0-\mathbf{w}^*)\right\|_{\mathbf{H}}^2 + \frac{\beta}{1200}\cdot\|\mathbf{w}_0-\mathbf{w}^*\|_{\mathbf{H}_{k^\dagger:\infty}}^2\cdot\frac{d_{\texttt{eff}}}{K},$$

and

$$\texttt{VarianceError} \geq \frac{\sigma^2}{400}\cdot\frac{d_{\texttt{eff}}}{K}.$$

Here $k^* := \max\{k : \lambda_k \geq 1/(\gamma_0 K)\}$, $k^\dagger := \max\{k : \lambda_k \geq 1/(\gamma_0(s+K))\}$, and the effective dimension *is defined by*

$$d_{\texttt{eff}} := k^* + \gamma_0 K \sum_{k^*<i\leq k^\dagger}\lambda_i + \gamma_0^2 K(s+K)\sum_{i>k^\dagger}\lambda_i^2.$$

Theorem 4.2 provides a problem-dependent lower bound for last iterate SGD in the well-specified linear regression model. It shows that our variance error bound is tight up to constant; however, for our bias error bound, there is a gap ($1/(\gamma_0 K)$ vs. $(\mathbf{I}-\gamma_0\mathbf{H})^K\mathbf{H}_{0:k^*}$) between the upper and lower bounds in the first term, and is missing a factor of $\|\mathbf{w}_0-\mathbf{w}^*\|_{\mathbf{H}_{0:k^\dagger}}^2$ and a $\log(s+K)$ factor in the second term. These gaps are due to some technical difficulties to obtain an accurate bias bound on the last iterate of SGD. We leave it as a future work to close these gaps.

4.3. Comparison with Polynomially Decaying Stepsize

In terms of the statistical minimax rate, it is proved by Ge et al. (2019) that the last iterate of SGD performs better with geometrically decaying stepsize than with polynomially decaying stepsize. Nonetheless, their comparison is in terms of the *worst-case* performance, and Ge et al. (2019) did not rule out the possibility that there could exist some linear regression problems such that SGD generalizes better with polynomially decaying stepsize. Thanks to our sharp problem-dependent bounds on SGD with (tail) geometrically decaying stepsize, we are able to compare its performance with that of SGD with (tail) polynomially decaying stepsize, in an *instance-wise* manner. The *(tail) polynomially decaying stepsize* is formally defined by

$$\gamma_t = \begin{cases}\gamma_0, & 0 \leq t \leq s;\\ \gamma_0/(t-s)^a, & s < t \leq N,\end{cases} \tag{4}$$

for some $a \in [0,1]$. We then present a problem-dependent excess risk lower bound for the last iterate of SGD with stepsize scheme (4). Due to the space limit, the following theorem focuses on $a \in [0,1)$; the full version for $a \in [0,1]$ is stated as Theorem E.3 in Appendix E.

Theorem 4.3 (A lower bound for poly-decaying stepsizes)**.** *Consider last iterate SGD with stepsize scheme* (4)*. Suppose Assumptions 3.1, 3.2B and 3.3' hold. Suppose $\gamma_0 < 1/(4\lambda_1)$, $s\gamma_0 \geq \sum_{t>s}\gamma_t$, and $a \in [0,1)$. Then we have*

$$\mathbb{E}[L(\mathbf{w}_N) - L(\mathbf{w}^*)] = \frac{1}{2}\texttt{BiasError} + \frac{1}{2}\texttt{VarianceError},$$

where

$$\texttt{BiasError} \gtrsim \left\|(\mathbf{I}-\gamma_0\mathbf{H})^{s+\frac{2N^{1-a}}{1-a}}\cdot(\mathbf{w}_0-\mathbf{w}^*)\right\|_{\mathbf{H}}^2 + \beta\cdot\|\mathbf{w}_0-\mathbf{w}^*\|_{\mathbf{H}_{k^\dagger:\infty}}^2\cdot\frac{d_{\texttt{eff}}}{N},$$

and

$$\texttt{VarianceError} \gtrsim \sigma^2\cdot\frac{d_{\texttt{eff}}}{N}.$$

Here $k^* := \max\{k : \gamma_0\lambda_k \geq (1-a)/(2(N-s)^{1-a})\}$, $k^\dagger := \max\{k : \gamma_0\lambda_k \geq 1/(2s)\}$, and the effective dimension *is defined by*

$$d_{\texttt{eff}} := \sum_{i\leq k^*}\max\{N^{1-a}\gamma_0\lambda_i,\ a\log(N)\} + \gamma_0 N\sum_{k^*<i\leq k^\dagger}\lambda_i + \gamma_0^2 sN\sum_{i>k^\dagger}\lambda_i^2.$$

Comparing Theorem 4.3 for tail polynomially decaying stepsize with Theorem 4.1 for tail geometrically decaying stepsize, the main difference is in the definition of the effective dimension $d_{\texttt{eff}}$. This is due to the different algorithmic regularization effects afforded by the different stepsize decaying schemes. With this difference in hand, our next theorem provides an instant-wise *risk inflation* (Dhillon et al., 2013) comparison between (the last iterate of SGD with) these two stepsize decaying schemes.

Theorem 4.4 (An instance-wise risk comparison)**.** *Suppose Assumptions 3.1, 3.2 and 3.3' all hold. Suppose $\gamma_0 <$*

$1/(3\alpha\,\mathrm{tr}(\mathbf{H})\log(s+K))$. *Let N be the sample size, and set $s = N/2$. Let $\mathbf{w}_N^{\exp}$ and $\mathbf{w}_N^{\mathrm{poly}}$ be the last iterate of SGD with stepsize scheme (3) and (4), respectively. Then there is a constant $C > 0$ such that*

$$\mathbb{E}[L(\mathbf{w}_N^{\exp}) - L(\mathbf{w}^*)] \le C \cdot \big(1 + \log(N) \cdot R(N)\big) \cdot \mathbb{E}[L(\mathbf{w}_N^{\mathrm{poly}}) - L(\mathbf{w}^*)]$$

for every problem-algorithm instance $(\mathbf{H}, \mathbf{w}^*, \gamma_0)$. Here

$$R(N) := \frac{\|\mathbf{w} - \mathbf{w}^*\|^2_{\mathbf{I}_{0:k^\dagger}}/(\gamma_0 N) + \|\mathbf{w} - \mathbf{w}^*\|^2_{\mathbf{H}_{k^\dagger:\infty}}}{\sigma^2}$$

for $k^\dagger := \max\{k : \lambda_k \ge 1/(\gamma_0 N)\}$.

The choice of $s = N/2$ in Theorem 4.4 is for ensuring that the two SGD variants have the same optimization trajectory length, i.e., $\sum_{i=1}^N \gamma_i = \Theta(N\gamma_0)$. This rules out the trivial optimization difference in the bias error between the two SGD variants, so Theorem 4.4 reveals only the statistical difference between the two stepsize schemes.

Let us assume $\log(N) \cdot R(N) \le 1$ for now. Then Theorem 4.4 reads that, *for every problem instance*, with the same initial stepsize, the excess risk of SGD with tail geometrically decaying stepsize is *no worse than* that of SGD with tail polynomially decaying stepsize, *upto constant factors*. This suggests that for the last iterate of SGD, a tail geometrically decaying stepsize is *always* as good as a tail polynomially decaying stepsize in terms of generalization.

We now discuss the quantity $\log(N) \cdot R(N)$ in Theorem 4.4. First of all, this quantity is rooted from the $\log(s+K)$ factor in the bias error upper bound in Theorem 4.1. Therefore, the $\log(N) \cdot R(N)$ factor in Theorem 4.4 might be an artifact that can be removed given a tighter bias analysis (we conjecture that Theorem 4.1 is not tight with the $\log(s+K)$ factor). Moreover, we argue that $\log(N) \cdot R(N)$ itself is small in many scenarios so that the comparison in Theorem 4.4 is still meaningful. To see this, note that

$$R(N) \le \|\mathbf{w} - \mathbf{w}^*\|_2^2/(\gamma_0 N \sigma^2)$$

by the definition of $k^\dagger$. Thus, we have $\log(N) \cdot R(N) = \mathcal{O}(1)$ so long as $\|\mathbf{w}_0 - \mathbf{w}^*\|_2^2 = \mathcal{O}\left(\sigma^2\gamma_0 N/\log(N)\right)\big)$.

Figure 2 provides further empirical verification to our comparison of the two stepsize schemes for the last iterate of SGD. We see from Figure 2 that the last iterate of SGD generalizes significantly better with tail geometrically decaying stepsize than with tail polynomially decaying stepsize.

5. Overview of the Proof Techniques

We now sketch the proof of Theorem 4.1 and highlight the key proof techniques. A complete proof is deferred to Appendix C. For simplicity, let us denote $L := \log(N-s)$ and $K := (N-s)/L$, and assume they are integers.

Bias-Variance Decomposition. We follow the well-known operator viewpoint for analyzing SGD iterates (Bach & Moulines, 2013; Dieuleveut et al., 2017; Jain et al., 2017a;b; Neu & Rosasco, 2018; Ge et al., 2019; Zou et al., 2021b). In particular, the excess risk can be decomposed into *bias error* and *variance error* (see Lemma B.2 in the appendix):

$$\mathbb{E}[L(\mathbf{w}_N) - L(\mathbf{w}^*)] \le \langle \mathbf{H}, \mathbf{B}_N \rangle + \langle \mathbf{H}, \mathbf{C}_N \rangle,$$

where $\mathbf{B}_N$ and $\mathbf{C}_N$ refer to the last *bias iterate* and the last *variance iterate* in the matrix space, respectively. More precisely, they are recursively defined by[2]

$$\begin{cases} \mathbf{B}_t = (\mathcal{I} - \gamma_t \mathcal{T}_t) \circ \mathbf{B}_{t-1}, & t \ge 1; \\ \mathbf{B}_0 = (\mathbf{w}_0 - \mathbf{w}^*)(\mathbf{w}_0 - \mathbf{w}^*)^\top, \end{cases} \tag{5}$$

$$\begin{cases} \mathbf{C}_t = (\mathcal{I} - \gamma_t \mathcal{T}_t) \circ \mathbf{C}_{t-1} + \gamma_t^2 \mathbf{\Sigma}, & t \ge 1; \\ \mathbf{C}_0 = 0. \end{cases} \tag{6}$$

Here $\mathcal{I} := \mathbf{I} \otimes \mathbf{I}$, $\mathcal{M} := \mathbb{E}[\mathbf{x} \otimes \mathbf{x} \otimes \mathbf{x} \otimes \mathbf{x}]$ and $\mathcal{T}_t := \mathbf{H} \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{H} - \gamma_t \mathcal{M}$ are operators on symmetric matrices (see Appendix B for their precise definitions). One can verify that for symmetric matrix $\mathbf{A}$,

$$(\mathcal{I} - \gamma_t \mathcal{T}_t) \circ \mathbf{A} = \mathbb{E}[(\mathbf{I} - \gamma_t \mathbf{x}\mathbf{x}^\top)\mathbf{A}(\mathbf{I} - \gamma_t \mathbf{x}\mathbf{x}^\top)].$$

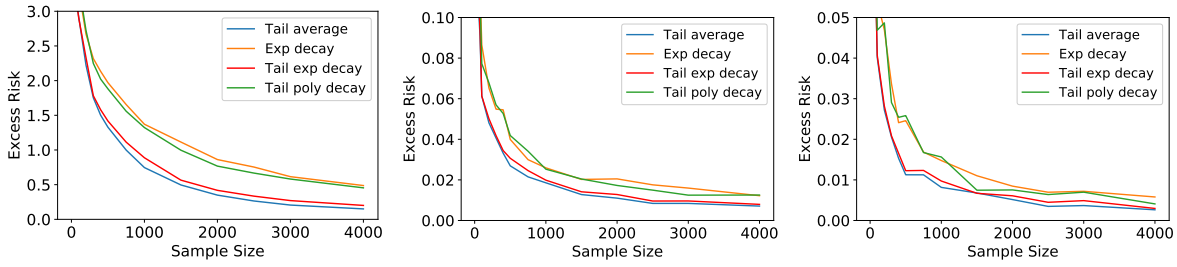
We next bound $\langle \mathbf{H}, \mathbf{B}_N \rangle$ and $\langle \mathbf{H}, \mathbf{C}_N \rangle$ separately.

5.1. Bias Upper Bound

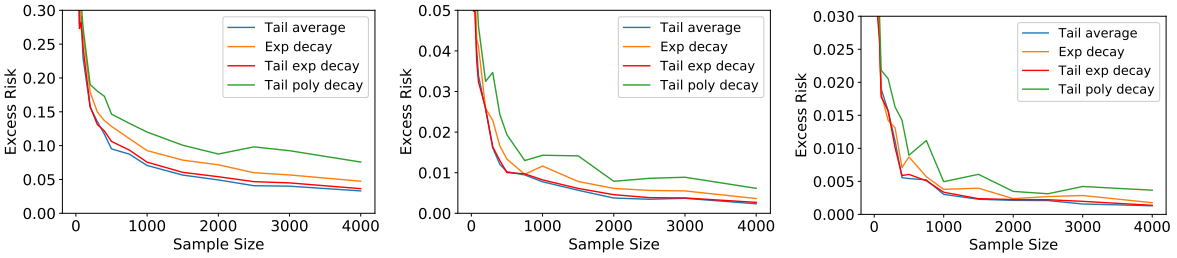
We first bound the bias error. Recall that the stepsize scheme (3) splits the total N iterations into $L = \log(N-s)$ fixed-stepsize phases: in the first phase, SGD is initialized from $\mathbf{B}_0$, and runs with constant stepsize γ_0 for $s+K$ steps; and in the ℓ-th phase for $2 \le \ell \le L$, SGD is initialized from $\mathbf{B}_{s+K(\ell-1)}$, and runs with stepsize $\gamma_0/2^{\ell-1}$ for K steps.

Main Challenges and Proof Techniques. The key difficulty here is to obtain a sharp characterization of *each bias iterate* (i.e., $\mathbf{B}_t$), instead of *their summation* $\sum_{t=1}^N \mathbf{B}_t$. Therefore, existing techniques for SGD with averaging (Jain et al., 2017b; Zou et al., 2021b) are not sufficient. In particular, Zou et al. (2021b) only gave a constant upper bound on the bias iterate (see Eq. (D.3) in their Lemma D.4, which is already sufficient for their purpose). To obtain a tight and vanishing bound on each bias iterate, we need to carefully utilize the $(\mathcal{I} - \gamma\widetilde{T})^i$ decaying factor in the bias expansion (see (8)). Our proof is motivated by this idea and handle the decaying factor with an inequality $(1-\gamma x)^t \le 1/(\gamma x)$ (see (9)). Based on this we get a (relatively loose) vanishing

[2] One can think of the bias iterates as SGD iterates on the data without additive label noise, and the variance iterates as SGD iterates with initialization $\mathbf{w}^*$.



(a) $\lambda_i = i^{-1}, \mathbf{w}^*[i] = 1$ (b) $\lambda_i = i^{-1}, \mathbf{w}^*[i] = i^{-1}$ (c) $\lambda_i = i^{-1}, \mathbf{w}^*[i] = i^{-2}$



(d) $\lambda_i = i^{-2}, \mathbf{w}^*[i] = 1$ (e) $\lambda_i = i^{-2}, \mathbf{w}^*[i] = i^{-1}$ (f) $\lambda_i = i^{-2}, \mathbf{w}^*[i] = i^{-2}$

Figure 2. Excess risk comparison between SGD variants. The problem dimension is $d = 256$ and the linear regression model is well-specified with noise variance $\sigma^2 = 1$. TAIL AVERAGE: constant-stepsize SGD with tail averaging ($s = N/2$); EXP DECAY: SGD with geometrically decaying stepsize ($s = 0$ and $K = \lceil N/\log(N) \rceil$); TAIL EXP DECAY: SGD with tail geometrically decaying stepsize ($s = N/2$ and $K = \lceil N/(2\log(N/2)) \rceil$); TAIL POLY DECAY: SGD with tail polynomially decaying stepsize ($s = N/2$ and $a = 1$). We consider 6 combinations of 2 different covariance matrices and 3 different true model parameters. For each algorithm and each sample size, we do a grid search and report the best excess risk achieved by $\gamma_0 \in \{10^{-4}, 2\times10^{-4}, 5\times10^{-4}, 7\times10^{-4}, 10^{-3}, 2\times10^{-3}, 5\times10^{-3}, 0.01, 0.02, 0.03, 0.05, 0.075, 0.1, 0.2, 0.3, 0.5, 0.8, 1.0\}$. The plots are averaged over 20 independent runs.

bound on each $\mathbf{B}_t$. We further sharpen this upper bound with a multi-phase strategy: (1) splitting the entire bias iterates into multiple phases; (2) deriving an upper bound for each phase; and (3) carefully combining them to get the final result. Details are explained below.

One Phase Analysis. We first investigate the decreasing effect of the bias error within one phase. For simplicity, with a slight abuse of notation, we use γ, n and $\mathbf{B}_t$ to denote the constant stepsize, the number of steps and the t-th bias iterate ($0 \le t \le n$) within one phase, respectively. Assume that $\gamma < 1/(3\alpha \operatorname{tr}(\mathbf{H}) \log n)$. Clearly $\mathbf{H}\otimes\mathbf{H} \succeq 0$, therefore

$$\begin{aligned}\widetilde{\mathcal{T}} &:= \mathbf{H}\otimes\mathbf{I} + \mathbf{I}\otimes\mathbf{H} - \gamma\mathbf{H}\otimes\mathbf{H} \\ &= \mathcal{T} + \gamma\mathcal{M} - \gamma\mathbf{H}\otimes\mathbf{H} \le \mathcal{T} + \gamma\mathcal{M}. \end{aligned} \tag{7}$$

Plug (7) into (5), and apply Assumption 3.2A, we obtain:

$$\mathbf{B}_t \preceq (\mathcal{I} - \gamma\widetilde{\mathcal{T}}) \circ \mathbf{B}_{t-1} + \alpha\gamma^2 \langle \mathbf{H}, \mathbf{B}_{t-1} \rangle \mathbf{H}, \quad t \ge 1.$$

Solving this recursion yields

$$\mathbf{B}_t \preceq (\mathcal{I}-\gamma\widetilde{\mathcal{T}})^t \circ \mathbf{B}_0 + \alpha\gamma^2 \sum_{i=0}^{t-1} (\mathcal{I}-\gamma\widetilde{\mathcal{T}})^{t-1-i} \circ \mathbf{H} \cdot \langle \mathbf{H}, \mathbf{B}_i \rangle . \tag{8}$$

In (8), we apply

$$(\mathcal{I} - \gamma\widetilde{\mathcal{T}})^{t-1-i} \circ \mathbf{H} = (\mathbf{I} - \gamma\mathbf{H})^{2(t-1-i)}\mathbf{H} \preceq \frac{\mathbf{I}}{\gamma(t-i)}$$

and take the inner product with $\mathbf{H}$, so we have

$$\langle \mathbf{H}, \mathbf{B}_t \rangle \le \big\langle (\mathcal{I} - \gamma\widetilde{\mathcal{T}})^t \circ \mathbf{H}, \mathbf{B}_0 \big\rangle + \alpha\gamma \operatorname{tr}(\mathbf{H}) \underbrace{\sum_{i=0}^{t-1} \frac{\langle \mathbf{H}, \mathbf{B}_i \rangle}{t-i}}_{(\diamond)}. \tag{9}$$

By recursively calling (9), one can observe that term $(\diamond)$ is *self-governed* (this trick first appears in Varre et al. (2021) to our knowledge), which leads to the following upper bound (see Lemma C.4 in the appendix):

$$(\diamond) \lesssim \Big\langle \sum_{i=0}^{t-1} \frac{(\mathcal{I} - \gamma\widetilde{\mathcal{T}})^i \circ \mathbf{H}}{t-i}, \mathbf{B}_0 \Big\rangle.$$

Substituting the above bound into (9) leads to

$$\begin{aligned}\langle \mathbf{H}, \mathbf{B}_t \rangle &\lesssim \Big\langle (\mathcal{I} - \gamma\widetilde{\mathcal{T}})^t \circ \mathbf{H} + \sum_{i=0}^{t-1} \frac{(\mathcal{I} - \gamma\widetilde{\mathcal{T}})^i \circ \mathbf{H}}{t-i}, \mathbf{B}_0 \Big\rangle \\ &\lesssim \Big\langle \frac{1}{\gamma t}\mathbf{I}_{0:k^*} + \mathbf{H}_{k^*:\infty}, \mathbf{B}_0 \Big\rangle, \end{aligned} \tag{10}$$

where the second inequality holds by bounding the summation $\sum_{0\le i<t}(\cdot)$ separately for $\sum_{0\le i<t/2}(\cdot)$ and $\sum_{t/2\le i<t}(\cdot)$ (see Lemma C.4 in the appendix for more details). Here k^* can be arbitrary. From (10), we can see a decreasing effect of the bias error within one phase.

Combining Multiple Phases. Next we discuss how to combine the decreasing effect of multiple phases. In this part, we use $\mathbf{B}^{(\ell)}$ to denote the bias iterate output by the ℓ-th phase (a.k.a., the input of the $(\ell+1)$-th phase).

In the first phase, a bound on $\mathbf{B}^{(1)}$ is obtained by setting $k^* = k^\dagger$ and $\gamma = \gamma_0$ in (10), and substituting (10) into (8) with $t = s + K$ (see Lemma C.7 in the appendix):

$$\mathbf{B}^{(1)} \lesssim (\mathcal{I} - \gamma\widetilde{\mathcal{T}})^{s+K} \circ \mathbf{B}_0 + \gamma_0^2(s+K)\log(s+K)\cdot \Big\langle \frac{\mathbf{I}_{0:k^\dagger}}{\gamma_0(s+K)} + \mathbf{H}_{k^\dagger:\infty}, \mathbf{B}_0 \Big\rangle \cdot \Big(\frac{\mathbf{I}_{0:k^\dagger}}{\gamma_0(s+K)} + \mathbf{H}_{k^\dagger:\infty}\Big). \tag{11}$$

In the second phase, setting $\gamma = \gamma_0/2$ and $t = K$ in (10), we obtain

$$\langle \mathbf{H}, \mathbf{B}^{(2)} \rangle \lesssim \Big\langle \frac{1}{\gamma t}\mathbf{I}_{0:k^*} + \mathbf{H}_{k^*:\infty},\ \mathbf{B}^{(1)} \Big\rangle. \tag{12}$$

Plugging (11) into (12) shows that $\mathbf{B}^{(2)}$ already achieves the desired bias bound in Theorem 4.1. The remaining effort is to combine the effect from the third to the L-th phase, which leads to (see Lemma C.8 in the appendix):

$$\langle \mathbf{H}, \mathbf{B}^{(L)} \rangle \le e \cdot \langle \mathbf{H}, \mathbf{B}^{(2)} \rangle.$$

This completes the proof for the bias error.

5.2. Variance Upper Bound

Main Challenges and Proof Techniques. Note that we are considering the variance error of the last SGD iterate, thus we cannot utilize the effect of iterate averaging to decrease the variance error (Bach & Moulines, 2013; Jain et al., 2017a;b; Zou et al., 2021b). Instead, to achieve a vanishing variance bound on the last iterate, we need to consider the effect of stepsize decaying. More details are provided below.

We first observe a uniform but crude upper bound on the variance iterates (see Lemma C.1 in the appendix, and also Lemma 5 in Ge et al. (2019)):

$$\mathbf{C}_t \preceq \frac{\gamma_0\sigma^2}{1 - \alpha\gamma_0\,\mathrm{tr}(\mathbf{H})}\mathbf{I}, \quad t = 1, 2, \ldots, N. \tag{13}$$

Then we will plug this crude bound on $\mathbf{C}_t$ into (6) to further improve the upper bound of $\mathbf{C}_t$ (see Theorem C.2 and its proof in the appendix):

$$\mathbf{C}_N \preceq \frac{\sigma^2}{1 - \gamma_0 R^2} \underbrace{\sum_{t=1}^{N} \gamma_t^2 \prod_{i=t+1}^{N} (\mathbf{I} - \gamma_i\mathbf{H})^2\mathbf{H}}_{(*)}.$$

The remaining effort is to control term $(*)$. Intuitively, though $(*)$ is a summation of N terms, $(*)$ could vanish as N increases thanks to the appropriate decaying stepsize scheme (3): for large t, the t-th term in the summation is small as γ_t is small; as for small t where γ_t is large, the t-th term in the summation is also small since the product $\prod_{i=t+1}^{N}(\mathbf{I} - \gamma_i\mathbf{H})^2\mathbf{H}$ is small (note the subsequent γ_i's are at least $\gamma_t/2$ according to (3)). More precisely, our analysis (see Lemmas C.3 and D.2 in the appendix) shows that, ignoring constant factors,

$$(*) \eqsim \frac{1}{K}\mathbf{H}_{0:k^*}^{-1} + \gamma_0\mathbf{I}_{k^*:k^\dagger} + \gamma_0^2(s+K)\mathbf{H}_{k^\dagger:\infty}$$

for the optimally chosen k^* and $k^\dagger$ in Theorems 4.1 and 4.2. In this way, we can establish a tight upper bound on $\mathbf{C}_N$. Finally, taking inner product with $\mathbf{H}$ yields the variance upper bound (see Theorem C.2 in the appendix).

6. Concluding Remarks

In this work, we provide a problem dependent excess risk bound for the last iterate of SGD with decaying stepsize for linear regression. The derived bound is dimension-free and can be applied to the overparamerized setting where the problem dimension excesses the sample size. A nearly-matching, problem-dependent lower bound is also proved. We further compare the excess risk bounds of last iterate SGD with tail geometric-decaying stepsize and that with tail polynomial-decaying stepsize, and show that the former outperforms the latter, instance-wisely. We believe the developed theoretical framework can also be used to find better stepsize schemes, or even the optimal one, which is left as a future work.

Acknowledgements

We would like to thank the anonymous reviewers and area chairs for their helpful comments. JW and VB are supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR00112190130. DZ acknowledges the support from Bloomberg Data Science Ph.D. Fellowship. QG is partially supported by the National Science Foundation award IIS-1906169 and IIS-2008981. SK acknowledges funding from the Office of Naval Research under award N00014-22-1-2377 and the National Science Foundation Grant under award CCF-1703574. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

References

Aybat, N. S., Fallah, A., Gurbuzbalaban, M., and Ozdaglar, A. A universally optimal multistage accelerated stochastic gradient method. *Advances in neural information processing systems*, 32:8525–8536, 2019.

Bach, F. and Moulines, E. Non-strongly-convex smooth

stochastic approximation with convergence rate $o(1/n)$. *Advances in neural information processing systems*, 26: 773–781, 2013.

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 2020.

Belkin, M., Hsu, D., and Xu, J. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020.

Berthier, R., Bach, F., and Gaillard, P. Tight nonparametric convergence rates for stochastic gradient descent under the noiseless linear model. *arXiv preprint arXiv:2006.08212*, 2020.

Bubeck, S. Theory of convex optimization for machine learning. *arXiv preprint arXiv:1405.4980*, 15, 2014.

Davis, D., Drusvyatskiy, D., and Charisopoulos, V. Stochastic algorithms with geometric step decay converge linearly on sharp functions. *arXiv preprint arXiv:1907.09547*, 2019.

Dekel, O., Gilad-Bachrach, R., Shamir, O., and Xiao, L. Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research*, 13(1), 2012.

Dhillon, P. S., Foster, D. P., Kakade, S. M., and Ungar, L. H. A risk comparison of ordinary least squares vs ridge regression. *The Journal of Machine Learning Research*, 14(1):1505–1511, 2013.

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

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

Fang, K.-T., Kotz, S., and Ng, K. W. *Symmetric multivariate and related distributions*. Chapman and Hall/CRC, 2018.

Ge, R., Kakade, S. M., Kidambi, R., and Netrapalli, P. The step decay schedule: A near optimal, geometrically decaying learning rate procedure for least squares. *arXiv preprint arXiv:1904.12838*, 2019.

Ghadimi, S. and Lan, G. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012.

Hazan, E. and Kale, S. Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization. *The Journal of Machine Learning Research*, 15(1):2489–2512, 2014.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. corr abs/1512.03385 (2015), 2015.

Hsu, D. J., Kakade, S. M., and Zhang, T. Random design analysis of ridge regression. *Foundations of Computational Mathematics*, 14(3):569–600, 2014.

Jain, P., Kakade, S. M., Kidambi, R., Netrapalli, P., Pillutla, V. K., and Sidford, A. A markov chain theory approach to characterizing the minimax optimality of stochastic gradient descent (for least squares). *arXiv preprint arXiv:1710.09430*, 2017a.

Jain, P., Netrapalli, P., Kakade, S. M., Kidambi, R., and Sidford, A. Parallelizing stochastic gradient descent for least squares regression: mini-batching, averaging, and model misspecification. *The Journal of Machine Learning Research*, 18(1):8258–8299, 2017b.

Kulunchakov, A. and Mairal, J. A generic acceleration framework for stochastic composite optimization. *arXiv preprint arXiv:1906.01164*, 2019.

Lacoste-Julien, S., Schmidt, M., and Bach, F. A simpler approach to obtaining an o (1/t) convergence rate for the projected stochastic subgradient method. *arXiv preprint arXiv:1212.2002*, 2012.

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

Mücke, N., Neu, G., and Rosasco, L. Beating sgd saturation with tail-averaging and minibatching. *Advances in Neural Information Processing Systems*, 32, 2019.

Neu, G. and Rosasco, L. Iterate averaging as regularization for stochastic gradient descent. In *Conference On Learning Theory*, pp. 3222–3242. PMLR, 2018.

Pan, R., Ye, H., and Zhang, T. Eigencurve: Optimal learning rate schedule for sgd on quadratic objectives with skewed hessian spectrums. *arXiv preprint arXiv:2110.14109*, 2021.

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

Rakhlin, A., Shamir, O., and Sridharan, K. Making gradient descent optimal for strongly convex stochastic optimization. *arXiv preprint arXiv:1109.5647*, 2011.

Schölkopf, B., Smola, A. J., Bach, F., et al. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

Tsigler, A. and Bartlett, P. L. Benign overfitting in ridge regression. *arXiv preprint arXiv:2009.14286*, 2020.

Varre, A., Pillaud-Vivien, L., and Flammarion, N. Last iterate convergence of sgd for least-squares in the interpolation regime. *arXiv preprint arXiv:2102.03183*, 2021.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Zou, D., Wu, J., Braverman, V., Gu, Q., Foster, D. P., and Kakade, S. M. The benefits of implicit regularization from sgd in least squares problems. *The 35th Conference on Neural Information Processing Systems*, 2021a.

Zou, D., Wu, J., Braverman, V., Gu, Q., and Kakade, S. M. Benign overfitting of constant-stepsize sgd for linear regression. *The 34th Annual Conference on Learning Theory*, 2021b.

A. More Examples for Assumption 3.2

Proposition A.1 (Examples for Assumption 3.2A)**.** *Assumption 3.2A holds for data distributions with a* bounded kurtosis along every direction *(Dieuleveut et al., 2017), i.e., there is a constant $\alpha > 0$ such that*

$$\text{for every } \mathbf{v},\ \mathbb{E}[\langle \mathbf{v}, \mathbf{x}\rangle^4] \le \alpha \langle \mathbf{v}, \mathbf{H}\mathbf{v}\rangle^2 .$$

In particular, the above is satisfied when $\mathbf{z} := \mathbf{H}^{-\frac{1}{2}}\mathbf{x}$ has sub-Gaussian or sub-exponential tail.

Proof. For a PSD matrix $\mathbf{A}$, with eigenvalues $(\mu_i)_{i\ge 1}$ and eigenvectors $(\mathbf{v}_i)_{i\ge 1}$, and a vector $\mathbf{u}$, we have

$$\begin{aligned}
\mathbf{u}^\top \mathbb{E}[\mathbf{x}\mathbf{x}^\top \mathbf{A}\mathbf{x}\mathbf{x}^\top]\mathbf{u} &= \mathbb{E}[\mathbf{x}^\top \mathbf{A}\mathbf{x}\cdot \langle \mathbf{x}, \mathbf{u}\rangle^2] \\
&= \sum_i \mu_i \cdot \mathbb{E}[\langle \mathbf{x}, \mathbf{v}_i\rangle^2 \cdot \langle \mathbf{x}, \mathbf{u}\rangle^2] \\
&\le \sum_i \mu_i \cdot \sqrt{\mathbb{E}[\langle \mathbf{x}, \mathbf{v}_i\rangle^4]\cdot \mathbb{E}[\langle \mathbf{x}, \mathbf{u}\rangle^4]} \qquad \text{(by Cauchy-Schwarz inequality)} \\
&\le \alpha \cdot \sum_i \mu_i \cdot \langle \mathbf{v}_i, \mathbf{H}\mathbf{v}_i\rangle \cdot \langle \mathbf{u}, \mathbf{H}\mathbf{u}\rangle \qquad \text{(by bounded kurtosis condition)} \\
&= \alpha \cdot \langle \mathbf{A}, \mathbf{H}\rangle \cdot \langle \mathbf{u}, \mathbf{H}\mathbf{u}\rangle.
\end{aligned}$$

Since the above holds for every vector $\mathbf{u}$, we conclude that

$$\mathbb{E}[\mathbf{x}\mathbf{x}^\top \mathbf{A}\mathbf{x}\mathbf{x}^\top] \preceq \alpha \cdot \langle \mathbf{A}, \mathbf{H}\rangle \cdot \mathbf{H},$$

which proves Assumption 3.2A. □

Proposition A.2 (Examples for Assumption 3.2B)**.** *Denote $\mathbf{z} := \mathbf{H}^{-\frac{1}{2}}\mathbf{x} =: (z_1, \dots, z_d)^\top$. Then Assumption 3.2B holds if:*

1. *the distribution of $\mathbf{z}$ is spherically symmetric, with a stochastic representation $\mathbf{z} = r\cdot \mathbf{u}$ where r and $\mathbf{u}$ are independent, $r > 0$ and $\mathbf{u}$ obeys the uniform distribution on the unit sphere $\mathcal{S}^{d-1}$;*
2. *$\mathbb{E}[r^2] = d$ and $\mathbb{E}[r^4] \ge \beta \cdot d(d+2)$ for a constant $\beta \ge 0.5$.*

Proof. We refer the reader to Fang et al. (2018) for the moments calculation of spherically symmetric distributions.

Note that $\mathbb{E}[\mathbf{z}] = \mathbf{0}$ and $\mathbb{E}[\mathbf{z}\mathbf{z}^\top] = \mathbf{I}$. Let $\mathbf{e}_1$ be the first standard basis, then for *every* unit vector $\mathbf{a} = (a_1, \dots, a_d)^\top$,

$$\begin{aligned}
\mathbb{E}[(\mathbf{e}_1^\top \mathbf{z})^2 \cdot (\mathbf{a}^\top \mathbf{z})^2] &= \mathbb{E}[z_1^2 \cdot (\mathbf{a}^\top \mathbf{z})^2] \\
&= a_1^2 \cdot \mathbb{E}[z_1^4] + \sum_{j=2}^d a_j^2 \cdot \mathbb{E}[z_1^2 z_j^2] \\
&= a_1^2 \cdot \mathbb{E}[r^4]\cdot \frac{3}{d(d+2)} + \sum_{j=2}^d a_j^2 \cdot \mathbb{E}[r^4]\cdot \frac{1}{d(d+2)} \\
&\ge 3\beta \cdot a_1^2 + \beta \cdot \sum_{j=2}^d a_j^2 \ge a_1^2 + \beta \\
&= (\mathbf{e}_1^\top \mathbf{a})^2 + \beta.
\end{aligned}$$

By the spherical symmetricity the above condition is equivalent to:

$$\text{for every unit vectors } \mathbf{a} \text{ and } \mathbf{b}, \quad \mathbb{E}[(\mathbf{a}^\top \mathbf{z})^2 \cdot (\mathbf{b}^\top \mathbf{z})^2] \ge (\mathbf{a}^\top \mathbf{b})^2 + \beta.$$

Then for every PSD matrix $\mathbf{A}$, with eigen decomposition $\mathbf{A} = \sum_{i=1}^d \mu_i \mathbf{a}_i \mathbf{a}_i^\top$, and every unit vector $\mathbf{b}$, it holds that

$$\mathbf{b}^\top \mathbb{E}[\mathbf{z}\mathbf{z}^\top \mathbf{A}\mathbf{z}\mathbf{z}^\top]\mathbf{b} = \sum_{i=1}^d \mu_i \cdot \mathbb{E}[(\mathbf{a}_i^\top \mathbf{z})^2 \cdot (\mathbf{b}^\top \mathbf{z})^2]$$

$$\geq \sum_{i=1}^{d} \mu_i \cdot ((\mathbf{a}_i^\top \mathbf{b})^2 + \beta)$$
$$= \mathbf{b}^\top \mathbf{A}\mathbf{b} + \beta \cdot \mathrm{tr}(\mathbf{A}),$$

which implies that $\mathbb{E}[\mathbf{z}\mathbf{z}^\top \mathbf{A}\mathbf{z}\mathbf{z}^\top] \succeq \mathbf{A} + \beta \cdot \mathrm{tr}(\mathbf{A}) \cdot \mathbf{I}$ for every PSD matrix $\mathbf{A}$. Finally, applying $\mathbf{x} = \mathbf{H}^{\frac{1}{2}}\mathbf{z}$ we obtain

$$\begin{aligned}\mathbb{E}[\mathbf{x}\mathbf{x}^\top \mathbf{A}\mathbf{x}\mathbf{x}^\top] &= \mathbf{H}^{\frac{1}{2}}\mathbb{E}[\mathbf{z}\mathbf{z}^\top \mathbf{H}^{\frac{1}{2}}\mathbf{A}\mathbf{H}^{\frac{1}{2}}\mathbf{z}\mathbf{z}^\top]\mathbf{H}^{\frac{1}{2}} \\ &\succeq \mathbf{H}^{\frac{1}{2}}\big(\mathbf{H}^{\frac{1}{2}}\mathbf{A}\mathbf{H}^{\frac{1}{2}} + \beta\,\mathrm{tr}(\mathbf{H}^{\frac{1}{2}}\mathbf{A}\mathbf{H}^{\frac{1}{2}})\mathbf{I}\big)\mathbf{H}^{\frac{1}{2}} \\ &= \mathbf{H}\mathbf{A}\mathbf{H} + \beta\,\mathrm{tr}(\mathbf{A}\mathbf{H})\mathbf{H},\end{aligned}$$

which proves Assumption 3.2B. □

B. Preliminaries

For two matrices $\mathbf{A} \in \mathbb{R}^{d\times d}$ and $\mathbf{B} \in \mathbb{R}^{d\times d}$, their tensor product is defined by

$$\mathbf{A} \otimes \mathbf{B} := \begin{pmatrix} a_{11}\mathbf{B} & \dots & \mathbf{a}_{1n}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{n1}\mathbf{B} & \dots & a_{nn}\mathbf{B} \end{pmatrix} \in \mathbb{R}^{d^2\times d^2},$$

where a_{ij} is the entry of $\mathbf{A}$ in the i-th row and j-th column. We can also understand $\mathbf{A} \otimes \mathbf{B}$ as a *linear matrix operator*, in which case we write

$$(\mathbf{A} \otimes \mathbf{B}) \circ \mathbf{C} := (\mathbf{A} \otimes \mathbf{B}) \cdot \mathtt{vector}(\mathbf{C}),$$

where $\mathbf{C} \in \mathbb{R}^{d\times d}$ is a matrix and $\mathtt{vector}(\mathbf{C}) \in \mathbb{R}^{d^2}$ converts $\mathbf{C}$ into a vector in the canonical manner.

Operators. We first summarize the linear operators (on symmetric matrices) to be used in the proof:

$$\mathcal{I} := \mathbf{I} \otimes \mathbf{I}, \qquad \mathcal{M} := \mathbb{E}[(\mathbf{x}\mathbf{x}^\top) \otimes (\mathbf{x}\mathbf{x}^\top)], \qquad \widetilde{\mathcal{M}} = \mathbf{H} \otimes \mathbf{H},$$
$$\mathcal{T}_t := \mathbf{H} \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{H} - \gamma_t \mathcal{M}, \qquad \widetilde{\mathcal{T}}_t = \mathbf{H} \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{H} - \gamma_t \mathbf{H} \otimes \mathbf{H}.$$

With a slight abuse of notations, we write $\mathcal{T}_t$ (resp. $\widetilde{\mathcal{T}}_t$) as $\mathcal{T}$ (resp. $\widetilde{\mathcal{T}}$) when the corresponding stepsize γ_t in its definition is written as γ. We use the notation $\mathcal{O} \circ \mathbf{A}$ to denotes the operator $\mathcal{O}$ acting on a symmetric matrix $\mathbf{A}$. One can verify the following rules for these operators acting on a symmetric matrix $\mathbf{A}$ (Zou et al., 2021b):

$$\begin{aligned}&\mathcal{I} \circ \mathbf{A} = \mathbf{A}, \qquad \mathcal{M} \circ \mathbf{A} = \mathbb{E}[(\mathbf{x}^\top \mathbf{A}\mathbf{x})\mathbf{x}\mathbf{x}^\top], \qquad \widetilde{\mathcal{M}} \circ \mathbf{A} = \mathbf{H}\mathbf{A}\mathbf{H}, \\ &(\mathcal{I} - \gamma\mathcal{T}) \circ \mathbf{A} = \mathbb{E}[(\mathbf{I} - \gamma\mathbf{x}\mathbf{x}^\top)\mathbf{A}(\mathbf{I} - \gamma\mathbf{x}\mathbf{x}^\top)], \quad (\mathcal{I} - \gamma\widetilde{\mathcal{T}}) \circ \mathbf{A} = (\mathbf{I} - \gamma\mathbf{H})\mathbf{A}(\mathbf{I} - \gamma\mathbf{H}).\end{aligned} \tag{14}$$

For the linear operators we have the following technical lemma from Zou et al. (2021b).

Lemma B.1 (Lemma B.1, Zou et al. (2021b))**.** *An operator $\mathcal{O}$ defined on symmetric matrices is called PSD mapping, if $\mathbf{A} \succeq 0$ implies $\mathcal{O} \circ \mathbf{A} \succeq 0$. Then we have*

1. *$\mathcal{M}$ and $\widetilde{\mathcal{M}}$ are both PSD mappings.*
2. *$\mathcal{I} - \gamma\mathcal{T}$ and $\mathcal{I} - \gamma\widetilde{\mathcal{T}}$ are both PSD mappings.*
3. *$\mathcal{M} - \widetilde{\mathcal{M}}$ and $\widetilde{\mathcal{T}} - \mathcal{T}$ are both PSD mappings.*
4. *If $0 < \gamma < 1/\lambda_1$, then $\widetilde{\mathcal{T}}^{-1}$ exists, and is a PSD mapping.*
5. *If $0 < \gamma < 1/(\alpha\,\mathrm{tr}(\mathbf{H}))$, then $\mathcal{T}^{-1} \circ \mathbf{A}$ exists for PSD matrix $\mathbf{A}$, and $\mathcal{T}^{-1}$ is a PSD mapping.*

Proof. See proof of Lemma B.1 in Zou et al. (2021b). □

We then prove the bias-variance decomposition.

Lemma B.2 (Bias-variance decomposition)**.** *Suppose Assumptions 3.1 and 3.3 hold. Then the excess risk could be decomposed as*

$$\mathbb{E}[L(\mathbf{w}_N) - L(\mathbf{w}^*)] \le \big\langle \mathbf{H}, \mathbf{B}_N \big\rangle + \big\langle \mathbf{H}, \mathbf{C}_N \big\rangle.$$

Proof. The proof has appeared in prior works (Jain et al., 2017b; Ge et al., 2019). For completeness, we provide a simplified (but not fully rigorous) proof here. Consider the centered SGD iterates $\boldsymbol{\eta}_t := \mathbf{w}_t - \mathbf{w}^*$, where $\mathbf{w}_t$ is given by (2), then the centered iterates are updated by

$$\boldsymbol{\eta}_t = (\mathbf{I} - \gamma_t \mathbf{x}_t \mathbf{x}_t^\top)\boldsymbol{\eta}_{t-1} + \gamma_t \xi_t \mathbf{x}_t, \quad t = 1, 2, \dots, N,$$

where $\xi_t := y_t - \langle \mathbf{w}^* - \mathbf{x}_t \rangle$ is the additive noise. With a slight abuse of probability spaces, one can view the centered SGD iterates as the sum of two random processes,

$$\boldsymbol{\eta}_t = \boldsymbol{\eta}_t^{\text{bias}} + \boldsymbol{\eta}_t^{\text{variance}}, \quad t = 1, 2, \dots, N, \tag{15}$$

where

$$\begin{cases} \boldsymbol{\eta}_t^{\text{bias}} = (\mathbf{I} - \gamma_t \mathbf{x}_t \mathbf{x}_t^\top)\boldsymbol{\eta}_{t-1}^{\text{bias}}; \\ \boldsymbol{\eta}_0^{\text{bias}} = \mathbf{w}_0 - \mathbf{w}^*, \end{cases} \qquad \begin{cases} \boldsymbol{\eta}_t^{\text{variance}} = (\mathbf{I} - \gamma_t \mathbf{x}_t \mathbf{x}_t^\top)\boldsymbol{\eta}_{t-1}^{\text{variance}} + \gamma_t \xi_t \mathbf{x}_t; \\ \boldsymbol{\eta}_0^{\text{variance}} = 0. \end{cases}$$

Then one can verify that $\mathbb{E}[\boldsymbol{\eta}_t^{\text{variance}}] = 0$, and moreover,

$$\mathbf{B}_t = \mathbb{E}[\boldsymbol{\eta}_t^{\text{bias}} \otimes \boldsymbol{\eta}_t^{\text{bias}}], \quad \mathbf{C}_t = \mathbb{E}[\boldsymbol{\eta}_t^{\text{variance}} \otimes \boldsymbol{\eta}_t^{\text{variance}}],$$

where $\mathbf{B}_t$ and $\mathbf{C}_t$ are defined in (5) and 6. Finally, the lemma is proved by

$$\begin{aligned} \mathbb{E}[L(\mathbf{w}_N) - L(\mathbf{w}^*)] &= \frac{1}{2} \langle \mathbf{H}, \mathbb{E}[\boldsymbol{\eta}_N \otimes \boldsymbol{\eta}_N] \rangle \\ &= \frac{1}{2} \big\langle \mathbf{H}, \mathbb{E}[(\boldsymbol{\eta}_N^{\text{bias}} + \boldsymbol{\eta}_N^{\text{variance}}) \otimes (\boldsymbol{\eta}_N^{\text{bias}} + \boldsymbol{\eta}_N^{\text{variance}})] \big\rangle \\ &\le \big\langle \mathbf{H}, \mathbb{E}[\boldsymbol{\eta}_N^{\text{bias}} \otimes \boldsymbol{\eta}_N^{\text{bias}}] \big\rangle + \big\langle \mathbf{H}, \mathbb{E}[\boldsymbol{\eta}_N^{\text{variance}} \otimes \boldsymbol{\eta}_N^{\text{variance}}] \big\rangle \\ &= \big\langle \mathbf{H}, \mathbf{B}_N \big\rangle + \big\langle \mathbf{H}, \mathbf{C}_N \big\rangle, \end{aligned}$$

where the inequality is because: for two vectors $\mathbf{u}$ and $\mathbf{v}$, $(\mathbf{u}+\mathbf{v})(\mathbf{u}+\mathbf{v})^\top \preceq 2(\mathbf{u}\mathbf{u}^\top + \mathbf{v}\mathbf{v}^\top)$. □

Lemma B.3 (Bias-variance decomposition, lower bound)**.** *Suppose Assumptions 3.1 and 3.3' hold. Then the excess risk could be decomposed as*

$$\mathbb{E}[L(\mathbf{w}_N) - L(\mathbf{w}^*)] = \frac{1}{2}\big\langle \mathbf{H}, \mathbb{E}[\boldsymbol{\eta}_N \otimes \boldsymbol{\eta}_N] \big\rangle = \frac{1}{2}\big\langle \mathbf{H}, \mathbf{B}_N \big\rangle + \frac{1}{2}\big\langle \mathbf{H}, \mathbf{C}_N \big\rangle.$$

Proof. The first equality is clear from definitions. The second equality is due to (15) and the following fact (by Assumption 3.3'): $\mathbb{E}[\boldsymbol{\eta}_t^{\text{variance}} | \boldsymbol{\eta}_t^{\text{bias}}] = 0$. □

C. Proof of Upper Bound

C.1. Variance Upper Bound

In this part we replace Assumption 3.2B with the following relaxed Assumption 3.2'. It is clear that Assumption 3.2B implies Assumption 3.2' with $R^2 = \alpha \operatorname{tr}(\mathbf{H})$.

Assumption 3.2' (Fourth moment condition, relaxed version)**.** *There exists a constant $R > 0$ such that $\mathbb{E}[\mathbf{x}\mathbf{x}^\top \mathbf{x}\mathbf{x}^\top] \preceq R^2 \mathbf{H}$.*

The following lemma is from Ge et al. (2019).

Lemma C.1 (Lemma 5 in Ge et al. (2019))**.** *Suppose Assumptions 3.1, 3.2' and 3.3 hold. Consider (6). Suppose $\gamma_0 < 1/R^2$. Then for every t we have*

$$\mathbf{C}_t \le \frac{\gamma_0 \sigma^2}{1 - \gamma_0 R^2} \mathbf{I}.$$

Proof. The original proof has appeared in Ge et al. (2019) and Jain et al. (2017a). We present a proof here for completeness. We proceed with induction. For $t=0$ we have $\mathbf{C}_0 = 0 \preceq \frac{\gamma_0\sigma^2}{1-\gamma_0 R^2}\mathbf{I}$. We then assume that $\mathbf{C}_{t-1} \preceq \frac{\gamma_0\sigma^2}{1-\gamma_0 R^2}\mathbf{I}$, and exam $\mathbf{C}_t$ based on (6):

$$\begin{aligned}
\mathbf{C}_t &= (\mathcal{I} - \gamma_t \mathcal{T}_t)\circ \mathbf{C}_{t-1} + \gamma_t^2 \mathbf{\Sigma} \\
&= (\mathcal{I} - \gamma_t \mathbf{H}\otimes\mathbf{I} - \gamma_t \mathbf{I}\otimes\mathbf{H})\circ \mathbf{C}_{t-1} + \gamma_t^2 \mathcal{M}\circ \mathbf{C}_{t-1} + \gamma_t^2 \mathbf{\Sigma} \\
&\preceq \frac{\gamma_0\sigma^2}{1-\gamma_0 R^2}\cdot(\mathbf{I} - 2\gamma_t \mathbf{H}) + \gamma_t^2 \cdot \frac{\gamma_0\sigma^2}{1-\gamma_0 R^2}\cdot R^2\mathbf{H} + \gamma_t^2\sigma^2\mathbf{H} \\
&= \frac{\gamma_0\sigma^2}{1-\gamma_0 R^2}\cdot\mathbf{I} - (2\gamma_t\gamma_0 - \gamma_t^2)\cdot\frac{\sigma^2}{1-\gamma_0 R^2}\mathbf{H} \\
&\preceq \frac{\gamma_0\sigma^2}{1-\gamma_0 R^2}\cdot\mathbf{I}.
\end{aligned}$$

This completes the induction. □

Theorem C.2 (A variance bound)**.** *Suppose Assumptions 3.1, 3.2' and 3.3 hold. Consider (6). Let $K = (N-s)/\log(N-s)$. Suppose $s \geq 0$, $K \geq 1$ and $\gamma_0 < 1/R^2$. We have*

$$\langle \mathbf{H}, \mathbf{C}_N\rangle \leq \frac{8\sigma^2}{1-\gamma_0 R^2}\left(\frac{k^*}{K} + \gamma_0 \sum_{k^*<i\leq k^\dagger}\lambda_i + \gamma_0^2(s+K)\sum_{i>k^\dagger}\lambda_i^2\right),$$

where k^* and $k^\dagger$ can be arbitrary.

Proof. From (6) we have

$$\begin{aligned}
\mathbf{C}_t &\preceq (\mathcal{I} - \gamma_t\widetilde{\mathcal{T}}_t)\circ\mathbf{C}_{t-1} + \gamma_t^2\mathcal{M}\circ\mathbf{C}_{t-1} + \gamma_t^2\sigma^2\mathbf{H} \\
&\preceq (\mathcal{I} - \gamma_t\widetilde{\mathcal{T}}_t)\circ\mathbf{C}_{t-1} + \gamma_t^2\cdot\frac{\gamma_0\sigma^2}{1-\gamma_0 R^2}\cdot R^2\mathbf{H} + \gamma_t^2\sigma^2\mathbf{H} \qquad \text{(use Lemma C.1)} \\
&= (\mathcal{I} - \gamma_t\widetilde{\mathcal{T}}_t)\circ\mathbf{C}_{t-1} + \frac{\gamma_t^2\sigma^2}{1-\gamma_0 R^2}\mathbf{H}.
\end{aligned}$$

Solving the recursion yields

$$\begin{aligned}
\mathbf{C}_N &\preceq \frac{\sigma^2}{1-\gamma_0 R^2}\sum_{t=1}^N \gamma_t^2 \prod_{i=t+1}^N (\mathcal{I} - \gamma_i\widetilde{\mathcal{T}}_i)\circ\mathbf{H} \\
&= \frac{\sigma^2}{1-\gamma_0 R^2}\sum_{t=1}^N \gamma_t^2 \prod_{i=t+1}^N (\mathbf{I} - \gamma_i\mathbf{H})^2\mathbf{H} \\
&\preceq \frac{\sigma^2}{1-\gamma_0 R^2}\underbrace{\sum_{t=1}^N \gamma_t^2 \prod_{i=t+1}^N (\mathbf{I} - \gamma_i\mathbf{H})\mathbf{H}}_{(*)}.
\end{aligned} \tag{16}$$

Now recalling (3), we have

$$\begin{aligned}
(*) &= \gamma_0^2 \sum_{i=1}^{s+K}\Big(\mathbf{I} - \gamma_0\mathbf{H}\Big)^{s+K-i}\prod_{j=1}^{L-1}\Big(\mathbf{I} - \frac{\gamma_0}{2^j}\mathbf{H}\Big)^K\mathbf{H} \\
&\quad + \sum_{\ell=1}^{L-1}\Big(\frac{\gamma_0}{2^\ell}\Big)^2\sum_{i=1}^{K}\Big(\mathbf{I} - \frac{\gamma_0}{2^\ell}\mathbf{H}\Big)^{K-i}\prod_{j=\ell+1}^{L-1}\Big(\mathbf{I} - \frac{\gamma_0}{2^j}\mathbf{H}\Big)^K\mathbf{H} \\
&= \gamma_0\Big(\mathbf{I} - \Big(\mathbf{I} - \gamma_0\mathbf{H}\Big)^{s+K}\Big)\prod_{j=1}^{L-1}\Big(\mathbf{I} - \frac{\gamma_0}{2^j}\mathbf{H}\Big)^K
\end{aligned}$$

$$+\sum_{\ell=1}^{L-1}\frac{\gamma_0}{2^\ell}\left(\mathbf{I}-\left(\mathbf{I}-\frac{\gamma_0}{2^\ell}\mathbf{H}\right)^K\right)\prod_{j=\ell+1}^{L-1}\left(\mathbf{I}-\frac{\gamma_0}{2^j}\mathbf{H}\right)^K, \tag{17}$$

where we understand $\prod_{j=L}^{L-1}(\cdot)=1$. Define a scalar function

$$f(x):=x\cdot\left(1-\left(1-x\right)^{s+K}\right)\cdot\prod_{j=1}^{L-1}\left(1-\frac{x}{2^j}\right)^K+\sum_{\ell=1}^{L-1}\frac{x}{2^\ell}\cdot\left(1-\left(1-\frac{x}{2^\ell}\right)^K\right)\cdot\prod_{j=\ell+1}^{L-1}\left(1-\frac{x}{2^j}\right)^K,$$

then applying $f(\cdot)$ to $\gamma_0\mathbf{H}$ in each diagonal entry, and using Lemma C.3, we have

$$f(\gamma_0\mathbf{H})\preceq\frac{8}{K}\mathbf{I}_{0:k^*}+2\gamma_0\mathbf{H}_{k^*:k^\dagger}+2\gamma_0^2(s+K)\mathbf{H}^2_{k^\dagger:\infty},$$

for arbitrary k^* and $k^\dagger$. Now using (16) and (17), we obtain

$$\mathbf{C}_N\preceq\frac{\sigma^2}{1-\gamma_0R^2}\cdot f(\gamma_0\mathbf{H})\cdot\mathbf{H}^{-1}\preceq\frac{8\sigma^2}{1-\gamma_0R^2}\left(\frac{1}{K}\mathbf{H}^{-1}_{0:k^*}+\gamma_0\mathbf{I}_{k^*:k^\dagger}+\gamma_0^2(s+K)\mathbf{H}_{k^*:\infty}\right),$$

and consequently,

$$\langle\mathbf{H},\mathbf{C}_N\rangle\le\frac{8\sigma^2}{1-\gamma_0R^2}\left(\frac{k^*}{K}+\gamma_0\sum_{k^*<i\le k^\dagger}\lambda_i+\gamma_0^2(s+K)\sum_{i>k^\dagger}\lambda_i^2\right),$$

where k^* and $k^\dagger$ can be arbitrary.

□

Lemma C.3. *Suppose $s\ge0, K\ge1$ and $x\in(0,1]$. For the scalar function*

$$f(x):=x\cdot\left(1-\left(1-x\right)^{s+K}\right)\cdot\prod_{j=1}^{L-1}\left(1-\frac{x}{2^j}\right)^K+\sum_{\ell=1}^{L-1}\frac{x}{2^\ell}\cdot\left(1-\left(1-\frac{x}{2^\ell}\right)^K\right)\cdot\prod_{j=\ell+1}^{L-1}\left(1-\frac{x}{2^j}\right)^K,$$

we have

$$f(x)\le\min\left\{2(s+K)x^2,\ 2x,\ \frac{8}{K}\right\}.$$

Proof. We show each upper bound separately.

- For $x\in(0,1]$, we have $(1-x)^{s+K}\ge1-(s+K)x$ and $(1-x)^K\ge1-Kx$, which lead to
$$f(x)\le x\cdot(s+K)x\cdot1+\sum_{\ell=1}^{L-1}\frac{x}{2^\ell}\cdot\frac{Kx}{2^\ell}\cdot1\le2(s+K)x^2.$$
- Clearly, for $x\in(0,1]$ we have: $f(x)\le x\cdot1\cdot1+\sum_{\ell=1}^{L-1}\frac{x}{2^\ell}\cdot1\cdot1\le2x$.
- For $x\in(0,2/K)$, by the previous bound we have $f(x)\le2x\le4/K$.
  As for $x\in[2/K,1]$, there is an
  $$\ell^*:=\lfloor\log(Kx)\rfloor-1\in[0,\ L-1),$$
  such that
  $$2^{\ell^*+1}/K\le x<2^{\ell^*+2}/K.$$
  by which and the definition of $f(x)$ we obtain:
  $$f(x)\le x\cdot1\cdot\prod_{j=1}^{L-1}\left(1-\frac{x}{2^j}\right)^K+\sum_{\ell=1}^{L-1}\frac{x}{2^\ell}\cdot1\cdot\prod_{j=\ell+1}^{L-1}\left(1-\frac{x}{2^j}\right)^K$$

$$
\begin{aligned}
&= \sum_{\ell=0}^{\ell^*} \frac{x}{2^\ell} \cdot \prod_{j=\ell+1}^{L-1} \big(1 - \frac{x}{2^j}\big)^K + \sum_{\ell=\ell^*+1}^{L-1} \frac{x}{2^\ell} \cdot \prod_{j=\ell+1}^{L-1} \big(1 - \frac{x}{2^j}\big)^K \\
&\le \sum_{\ell=0}^{\ell^*} \frac{x}{2^\ell} \cdot \big(1 - \frac{x}{2^{\ell+1}}\big)^K + \sum_{\ell=\ell^*+1}^{L-1} \frac{x}{2^\ell} \cdot 1 \\
&\le \sum_{\ell=0}^{\ell^*} \frac{2^{\ell^*-\ell+2}}{K} \cdot \Big(1 - \frac{2^{\ell^*-\ell}}{K}\Big)^K + \sum_{\ell=\ell^*+1}^{L-1} \frac{2^{\ell^*-\ell+2}}{K} \cdot 1 \\
&\le \frac{4}{K} \cdot \sum_{\ell=0}^{\ell^*} 2^{\ell^*-\ell} \cdot e^{-2^{\ell^*-\ell}} + \frac{4}{K} \\
&\le \frac{4}{K} \cdot 1 + \frac{4}{K} = \frac{8}{K}.
\end{aligned}
$$

In sum we have $f(x) \le 8/K$ holds for every $x \in (0, 1]$.

□

C.2. Preparation: Bias Upper Bound in a Single Phase

In this section we consider running bias iterates with constant stepsize γ for n steps. We note this process corresponds to SGD in one phase with constant stepsize. For simplicity we denote the initial bias iterate as $\mathbf{B}_0$. Then the bias iterates are updated according to

$$\mathbf{B}_t = (\mathcal{I} - \gamma\mathcal{T}) \circ \mathbf{B}_{t-1}, \quad t = 1, 2, \ldots, n. \tag{18}$$

For simplicity, let us define

$$\widehat{\mathbf{H}}_t := \frac{1}{\gamma t}\mathbf{I}_{0:k^*} + \mathbf{H}_{k^*:\infty}, \quad t \ge 1, \tag{19}$$

where $k^* \ge 0$ could be any integer.

From (18) we have

$$
\begin{aligned}
\mathbf{B}_t &\preceq (\mathcal{I} - \gamma\widetilde{\mathcal{T}}) \circ \mathbf{B}_{t-1} + \gamma^2 \mathcal{M} \circ \mathbf{B}_{t-1} \\
&\preceq (\mathcal{I} - \gamma\widetilde{\mathcal{T}})^t \circ \mathbf{B}_0 + \gamma^2 \sum_{i=0}^{t-1} (\mathcal{I} - \gamma\widetilde{\mathcal{T}})^{t-1-i} \circ \mathcal{M} \circ \mathbf{B}_i \\
&\preceq (\mathcal{I} - \gamma\widetilde{\mathcal{T}})^t \circ \mathbf{B}_0 + \alpha\gamma^2 \sum_{i=0}^{t-1} (\mathcal{I} - \gamma\widetilde{\mathcal{T}})^{t-1-i} \circ \mathbf{H} \cdot \langle \mathbf{H}, \mathbf{B}_i \rangle \\
&= (\mathcal{I} - \gamma\widetilde{\mathcal{T}})^t \circ \mathbf{B}_0 + \alpha\gamma^2 \sum_{i=0}^{t-1} (\mathbf{I} - \gamma\mathbf{H})^{2(t-1-i)}\mathbf{H} \cdot \langle \mathbf{H}, \mathbf{B}_i \rangle,
\end{aligned} \tag{20}
$$

where the inequality also holds for $t = 0$ with the understanding that $\sum_{i=0}^{-1} \cdot = 0$.

The following lemma provides a crude upper bound on $\langle \mathbf{H}, \mathbf{B}_n \rangle$.

Lemma C.4. *Suppose Assumptions 3.1 and 3.2 hold. Consider (18). Suppose $n \ge 1$ and $\gamma < 1/(2\alpha \operatorname{tr}(\mathbf{H}) \log n)$. We have*

$$\langle \mathbf{H}, \mathbf{B}_n \rangle \le \frac{2}{1 - 2\alpha\gamma \operatorname{tr}(\mathbf{H}) \log n} \cdot \left\langle \frac{1}{\gamma n}\mathbf{I}_{0:k^*} + \mathbf{H}_{k^*:\infty}, \mathbf{B}_0 \right\rangle,$$

where k^* can be arbitrary.

Proof. Notice $(1 - x)^t \le 1/((t + 1)x)$ for $x \in (0, 1)$, then $(\mathbf{I} - \gamma\mathbf{H})^{2t}\mathbf{H} \preceq \frac{1}{\gamma(t+1)}\mathbf{I}$. Inserting this into (20) and setting $t = n$, we obtain

$$\mathbf{B}_t \preceq (\mathcal{I} - \gamma\widetilde{\mathcal{T}})^t \circ \mathbf{B}_0 + \alpha\gamma \sum_{i=0}^{t-1} \frac{\langle \mathbf{H}, \mathbf{B}_t \rangle}{t - i} \cdot \mathbf{I}, \quad t \ge 1, \tag{21}$$

and thus

$$\begin{aligned}\langle \mathbf{H}, \mathbf{B}_t \rangle &\le \left\langle (\mathcal{I} - \gamma \widetilde{\mathcal{T}})^t \circ \mathbf{H}, \mathbf{B}_0 \right\rangle + \alpha\gamma \operatorname{tr}(\mathbf{H}) \sum_{i=0}^{t-1} \frac{\langle \mathbf{H}, \mathbf{B}_t \rangle}{t-i} \\ &= \left\langle (\mathbf{I} - \gamma \mathbf{H})^{2t} \mathbf{H}, \mathbf{B}_0 \right\rangle + \alpha\gamma \operatorname{tr}(\mathbf{H}) \sum_{i=0}^{t-1} \frac{\langle \mathbf{H}, \mathbf{B}_t \rangle}{t-i}, \quad n \ge 1. \end{aligned} \tag{22}$$

Recursively applying (22) to each $\langle \mathbf{H}, \mathbf{B}_t \rangle$, we obtain

$$\begin{aligned}\sum_{t=0}^{n-1} \frac{\langle \mathbf{H}, \mathbf{B}_t \rangle}{n-t} &\le \left\langle \sum_{t=0}^{n-1} \frac{(\mathbf{I} - \gamma \mathbf{H})^{2t} \mathbf{H}}{n-t}, \mathbf{B}_0 \right\rangle + \alpha\gamma \operatorname{tr}(\mathbf{H}) \sum_{t=0}^{n-1} \sum_{i=0}^{t-1} \frac{\langle \mathbf{H}, \mathbf{B}_i \rangle}{(n-t)(t-i)} \\ &= \left\langle \sum_{t=0}^{n-1} \frac{(\mathbf{I} - \gamma \mathbf{H})^{2t} \mathbf{H}}{n-t}, \mathbf{B}_0 \right\rangle + \alpha\gamma \operatorname{tr}(\mathbf{H}) \sum_{i=0}^{n-2} \frac{\langle \mathbf{H}, \mathbf{B}_i \rangle}{n-i} \sum_{t=i+1}^{n-1} \left(\frac{1}{n-t} + \frac{1}{t-i} \right) \\ &\le \left\langle \sum_{t=0}^{n-1} \frac{(\mathbf{I} - \gamma \mathbf{H})^{2t} \mathbf{H}}{n-t}, \mathbf{B}_0 \right\rangle + 2\alpha\gamma \operatorname{tr}(\mathbf{H}) \log n \cdot \sum_{i=0}^{n-1} \frac{\langle \mathbf{H}, \mathbf{B}_i \rangle}{n-i}, \end{aligned}$$

which implies that for $\gamma < 1/(2\alpha \operatorname{tr}(\mathbf{H}) \log n)$, we have

$$\sum_{t=0}^{n-1} \frac{\langle \mathbf{H}, \mathbf{B}_t \rangle}{n-t} \le \frac{1}{1 - 2\alpha\gamma \operatorname{tr}(\mathbf{H}) \log n} \cdot \Bigg\langle \underbrace{\sum_{t=0}^{n-1} \frac{(\mathbf{I} - \gamma \mathbf{H})^{2t} \mathbf{H}}{n-t}}_{(*)}, \mathbf{B}_0 \Bigg\rangle. \tag{23}$$

We would like to acknowledge Varre et al. (2021), from where we learn the trick to reach (23). Furthermore, we can bound $(*)$ as follows:

$$\begin{aligned}(*) &= \sum_{t=0}^{n/2-1} \frac{(\mathbf{I} - \gamma \mathbf{H})^{2t} \mathbf{H}}{n-t} + \sum_{t=n/2}^{n-1} \frac{(\mathbf{I} - \gamma \mathbf{H})^{2t} \mathbf{H}}{n-t} \\ &\le \frac{2}{n} \sum_{t=0}^{n/2-1} (\mathbf{I} - \gamma \mathbf{H})^{2t} \mathbf{H} + (\mathbf{I} - \gamma \mathbf{H})^{n} \mathbf{H} \sum_{t=n/2}^{n-1} \frac{1}{n-t} \\ &\le 2 \cdot \frac{\mathbf{I} - (\mathbf{I} - \gamma \mathbf{H})^n}{\gamma n} + \log n \cdot (\mathbf{I} - \gamma \mathbf{H})^n \mathbf{H} \\ &\le 2 \log n \cdot \left(\frac{\mathbf{I} - (\mathbf{I} - \gamma \mathbf{H})^n}{\gamma n} + (\mathbf{I} - \gamma \mathbf{H})^n \mathbf{H} \right). \end{aligned} \tag{24}$$

Finally, inserting (23) and (24) into (22), we obtain

$$\begin{aligned}\langle \mathbf{H}, \mathbf{B}_n \rangle &\le \left\langle (\mathbf{I} - \gamma \mathbf{H})^{2n} \mathbf{H}, \mathbf{B}_0 \right\rangle + \frac{2\alpha\gamma \operatorname{tr}(\mathbf{H}) \log n}{1 - 2\alpha\gamma \operatorname{tr}(\mathbf{H}) \log n} \left\langle \frac{\mathbf{I} - (\mathbf{I} - \gamma \mathbf{H})^n}{\gamma n} + (\mathbf{I} - \gamma \mathbf{H})^n \mathbf{H}, \mathbf{B}_0 \right\rangle \\ &\le \frac{1}{1 - 2\alpha\gamma \operatorname{tr}(\mathbf{H}) \log n} \cdot \left\langle \frac{\mathbf{I} - (\mathbf{I} - \gamma \mathbf{H})^n}{\gamma n} + (\mathbf{I} - \gamma \mathbf{H})^n \mathbf{H}, \mathbf{B}_0 \right\rangle \\ &\le \frac{2}{1 - 2\alpha\gamma \operatorname{tr}(\mathbf{H}) \log n} \cdot \left\langle \frac{1}{\gamma n} \mathbf{I}_{0:k^*} + \mathbf{H}_{k^*:\infty}, \mathbf{B}_0 \right\rangle, \end{aligned}$$

where the last inequality is because

$$\frac{\mathbf{I} - (\mathbf{I} - \gamma \mathbf{H})^n}{\gamma n} \preceq \frac{1}{\gamma n} \mathbf{I}_{0:k^*} + \mathbf{H}_{k^*:\infty}, \qquad (\mathbf{I} - \gamma \mathbf{H})^n \mathbf{H} \preceq \frac{1}{\gamma n} \mathbf{I}_{0:k^*} + \mathbf{H}_{k^*:\infty}. \tag{25}$$

□

The following lemma provides an upper bound for $\mathbf{B}_n$.

Lemma C.5. *Suppose Assumptions 3.1 and 3.2 hold. Consider (18). Suppose $n \geq 2$ and $\gamma < 1/(2\alpha \operatorname{tr}(\mathbf{H}) \log n)$. We have*

$$\mathbf{B}_n \preceq (\mathbf{I} - \gamma\mathbf{H})^n \cdot \mathbf{B}_0 \cdot (\mathbf{I} - \gamma\mathbf{H})^n + \frac{3\alpha\gamma^2 n \log n}{1 - 2\alpha\gamma \operatorname{tr}(\mathbf{H}) \log n} \cdot \langle \widehat{\mathbf{H}}_n, \mathbf{B}_0 \rangle \cdot \widehat{\mathbf{H}}_n.$$

Proof. We bring Lemma C.4 into (20) to obtain

$$\begin{aligned}
\mathbf{B}_n &\preceq (\mathcal{I} - \gamma\widetilde{\mathcal{T}})^n \circ \mathbf{B}_0 + \alpha\gamma^2 (\mathbf{I} - \gamma\mathbf{H})^{2(n-1)}\mathbf{H} \cdot \langle \mathbf{H}, \mathbf{B}_0 \rangle \\
&\quad + \alpha\gamma^2 \sum_{t=1}^{n-1} (\mathbf{I} - \gamma\mathbf{H})^{2(n-1-t)}\mathbf{H} \cdot \langle \mathbf{H}, \mathbf{B}_t \rangle \\
&\preceq (\mathcal{I} - \gamma\widetilde{\mathcal{T}})^n \circ \mathbf{B}_0 + \alpha\gamma^2 \underbrace{(\mathbf{I} - \gamma\mathbf{H})^{2(n-1)}\mathbf{H} \cdot \langle \mathbf{H}, \mathbf{B}_0 \rangle}_{(*)} \\
&\quad + \frac{2\alpha\gamma^2}{1 - 2\alpha\gamma \operatorname{tr}(\mathbf{H}) \log n} \cdot \underbrace{\sum_{t=1}^{n-1} (\mathbf{I} - \gamma\mathbf{H})^{2(n-1-t)}\mathbf{H} \cdot \langle \widehat{\mathbf{H}}_t, \mathbf{B}_0 \rangle}_{(**)},
\end{aligned} \tag{26}$$

where

$$\widehat{\mathbf{H}}_t := \frac{1}{\gamma t}\mathbf{I}_{0:k^*} + \mathbf{H}_{k^*:\infty}, \quad t \geq 1.$$

For term $(**)$, we bound it by

$$\begin{aligned}
(**) &= \sum_{t=1}^{n/2-1} (\mathbf{I} - \gamma\mathbf{H})^{2(n-1-t)}\mathbf{H} \cdot \langle \widehat{\mathbf{H}}_t, \mathbf{B}_0 \rangle + \sum_{t=n/2}^{n-1} (\mathbf{I} - \gamma\mathbf{H})^{2(n-1-t)}\mathbf{H} \cdot \langle \widehat{\mathbf{H}}_t, \mathbf{B}_0 \rangle \\
&\preceq \sum_{t=1}^{n/2-1} (\mathbf{I} - \gamma\mathbf{H})^{n}\mathbf{H} \cdot \langle \widehat{\mathbf{H}}_t, \mathbf{B}_0 \rangle + \sum_{t=n/2}^{n-1} (\mathbf{I} - \gamma\mathbf{H})^{n-1-t}\mathbf{H} \cdot \langle \widehat{\mathbf{H}}_{n/2}, \mathbf{B}_0 \rangle \\
&= (\mathbf{I} - \gamma\mathbf{H})^{n}\mathbf{H} \cdot \Big\langle \sum_{t=1}^{n/2-1} \widehat{\mathbf{H}}_t, \mathbf{B}_0 \Big\rangle + \frac{\mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^{n/2}}{\gamma} \cdot \langle \widehat{\mathbf{H}}_{n/2}, \mathbf{B}_0 \rangle \\
&\preceq (\mathbf{I} - \gamma\mathbf{H})^{n}\mathbf{H} \cdot \langle n(\log n - 1) \cdot \widehat{\mathbf{H}}_n, \mathbf{B}_0 \rangle + 2 \cdot \frac{\mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^{n}}{\gamma} \cdot \langle \widehat{\mathbf{H}}_n, \mathbf{B}_0 \rangle \\
&\preceq n(\log n - 1) \cdot \widehat{\mathbf{H}}_n \cdot \langle \widehat{\mathbf{H}}_n, \mathbf{B}_0 \rangle + 2 \cdot \widehat{\mathbf{H}}_n \cdot \langle \widehat{\mathbf{H}}_n, \mathbf{B}_0 \rangle \qquad \text{(use (25))} \\
&= (n \log n - n + 2) \cdot \langle \widehat{\mathbf{H}}_n, \mathbf{B}_0 \rangle \cdot \widehat{\mathbf{H}}_n.
\end{aligned}$$

In order to bound $(*)$, notice that for $n \geq 2$,

$$(\mathbf{I} - \gamma\mathbf{H})^{2(n-1)}\mathbf{H} \preceq \frac{1}{2\gamma(n-1)}\mathbf{I}_{0:k^*} + \mathbf{H}_{k^*:\infty} \preceq \widehat{\mathbf{H}}_n,$$

then we have

$$(*) \preceq \widehat{\mathbf{H}}_n \cdot \langle \mathbf{H}, \mathbf{B}_0 \rangle \preceq n \cdot \widehat{\mathbf{H}}_n \cdot \langle \widehat{\mathbf{H}}_n, \mathbf{B}_0 \rangle,$$

where the last inequality is because $\gamma < 1/\operatorname{tr}(\mathbf{H})$ implies $\lambda_1, \ldots, \lambda_{k^*} < 1/\gamma$ for every k^*.

Finally, bring the bounds on $(*)$ and $(**)$ into (26), we obtain

$$\begin{aligned}
\mathbf{B}_n &\preceq (\mathcal{I} - \gamma\widetilde{\mathcal{T}})^n \circ \mathbf{B}_0 + \alpha\gamma^2 \cdot n \cdot \widehat{\mathbf{H}}_n \cdot \langle \widehat{\mathbf{H}}_n, \mathbf{B}_0 \rangle \\
&\quad + \frac{2\alpha\gamma^2}{1 - 2\alpha\gamma \operatorname{tr}(\mathbf{H}) \log n} \cdot (n \log n - n + 2) \cdot \langle \widehat{\mathbf{H}}_n, \mathbf{B}_0 \rangle \cdot \widehat{\mathbf{H}}_n
\end{aligned}$$

$$\preceq (\mathcal{I}-\gamma\widetilde{\mathcal{T}})^n\circ\mathbf{B}_0+\frac{3\alpha\gamma^2 n\log n}{1-2\alpha\gamma\operatorname{tr}(\mathbf{H})\log n}\cdot\left\langle\widehat{\mathbf{H}}_n,\mathbf{B}_0\right\rangle\cdot\widehat{\mathbf{H}}_n.$$

Applying the definition of $\widetilde{\mathcal{T}}$ completes the proof.

□

C.3. Bias Upper Bound

Let us denote the bias iterate at the end of each stepsize-decaying phase by

$$\mathbf{B}^{(\ell)}:=\begin{cases}\mathbf{B}_0, & \ell=0;\\ \mathbf{B}_{s+K*\ell}, & \ell=1,2,\dots,L.\end{cases}\tag{27}$$

According to (3) and the above definition, we can interpret the SGD iterates (5) as follows: in phase $\ell=1$, SGD is initialized from $\mathbf{B}_0$ and runs for $s+K$ steps with constant stepsize $\gamma^{(1)}:=\gamma$, and output $\mathbf{B}^{(1)}$; in phase $\ell\geq 2$, SGD is initialized from $\mathbf{B}^{(\ell-1)}$ and runs for K steps with constant stepsize

$$\gamma^{(\ell-1)}:=\frac{\gamma_0}{2^{\ell-1}},$$

and output $\mathbf{B}^{(\ell)}$; the final output is $\mathbf{B}^{(L)}=\mathbf{B}_N$.

We now build an upper bound for bias error based on results obtained in Section C.2.

Lemma C.6. *Suppose Assumptions 3.1 and 3.2 hold. Consider (27) and (5). Suppose $\gamma_0<1/(3\alpha\operatorname{tr}(\mathbf{H})\log(s+K))$. We have*

$$\begin{aligned}&\text{for } \ell=1,\ \left\langle\mathbf{H},\mathbf{B}^{(1)}\right\rangle\leq 6\cdot\left\langle\frac{1}{\gamma_0(s+K)}\mathbf{I}_{0:k^*}+\mathbf{H}_{k^*:\infty},\ \mathbf{B}_0\right\rangle;\\ &\text{for } \ell\geq 2,\ \left\langle\mathbf{H},\mathbf{B}^{(\ell)}\right\rangle\leq 6\cdot\left\langle\frac{1}{\gamma^{(\ell-1)}K}\mathbf{I}_{0:k^*}+\mathbf{H}_{k^*:\infty},\ \mathbf{B}^{(\ell-1)}\right\rangle,\end{aligned}$$

where k^* can be arbitrary.

Proof. For $\left\langle\mathbf{H},\mathbf{B}^{(1)}\right\rangle$, we apply Lemma C.4 with $\gamma\to\gamma_0$ and $n\to s+K$, and use the condition that $\alpha\gamma\operatorname{tr}(\mathbf{H})\log(s+K)\leq 1/3$.

For $\left\langle\mathbf{H},\mathbf{B}^{(\ell)}\right\rangle$ with $\ell\geq 2$, we apply Lemma C.4 with $\gamma\to\gamma^{(\ell-1)}$, $n\to K$ and $\mathbf{B}_0\to\mathbf{B}^{(\ell-1)}$, and use the condition that $\alpha\gamma\operatorname{tr}(\mathbf{H})\log(K)\leq 1/3$.

□

Lemma C.7. *Suppose Assumptions 3.1 and 3.2 hold. Consider (27) and (5). Suppose $\gamma_0<1/(3\alpha\operatorname{tr}(\mathbf{H})\log(s+K))$. We have*

$$\begin{aligned}&\text{for } \ell=1,\ \mathbf{B}^{(1)}\preceq(\mathbf{I}-\gamma_0\mathbf{H})^{s+K}\cdot\mathbf{B}_0\cdot(\mathbf{I}-\gamma_0\mathbf{H})^{s+K}+\\ &\quad 9\alpha\gamma_0^2(s+K)\log(s+K)\cdot\left\langle\frac{1}{\gamma_0(s+K)}\mathbf{I}_{0:k^*}+\mathbf{H}_{k^*:\infty},\ \mathbf{B}_0\right\rangle\cdot\left(\frac{1}{\gamma_0(s+K)}\mathbf{I}_{0:k^\dagger}+\mathbf{H}_{k^\dagger:\infty}\right),\\ &\text{for } \ell\geq 2,\ \mathbf{B}^{(\ell)}\preceq(\mathbf{I}-\gamma^{(\ell-1)}\mathbf{H})^{K}\cdot\mathbf{B}^{(\ell-1)}\cdot(\mathbf{I}-\gamma^{(\ell-1)}\mathbf{H})^{K}+\\ &\quad 5\alpha(\gamma^{(\ell-1)})^2K\log K\cdot\left\langle\frac{1}{\gamma^{(\ell-1)}K}\mathbf{I}_{0:k^*}+\mathbf{H}_{k^*:\infty},\ \mathbf{B}^{(\ell-1)}\right\rangle\cdot\left(\frac{1}{\gamma^{(\ell-1)}K}\mathbf{I}_{0:k^\dagger}+\mathbf{H}_{k^\dagger:\infty}\right),\end{aligned}$$

where k^* and $k^\dagger$ can be arbitrary.

Proof. For $\mathbf{B}^{(1)}$, we apply Lemma C.5 with $\gamma\to\gamma_0$ and $n\to s+K$, and use the condition that $\alpha\gamma_0\operatorname{tr}(\mathbf{H})\log(s+K)\leq 1/3$.

For $\mathbf{B}^{(\ell)}$ with $\ell\geq 2$, we apply Lemma C.5 with $\gamma\to\gamma^{(\ell-1)}$, $n\to K$ and $\mathbf{B}_0\to\mathbf{B}^{(\ell-1)}$, and use the condition that $\alpha\gamma^{(\ell-1)}\operatorname{tr}(\mathbf{H})\log(K)\leq\alpha\gamma_0\operatorname{tr}(\mathbf{H})\log(s+K)/2\leq 1/6$. □

Lemma C.8. *Suppose Assumptions 3.1 and 3.2 hold. Consider (27) and (5). Suppose $\gamma_0 < 1/(3\alpha \operatorname{tr}(\mathbf{H}) \log(s+K))$. We have*

$$\big\langle \mathbf{H}, \mathbf{B}_N \big\rangle = \big\langle \mathbf{H}, \mathbf{B}^{(L)} \big\rangle \le e \cdot \big\langle \mathbf{H}, \mathbf{B}^{(2)} \big\rangle.$$

Proof. Let $\ell \ge 2$. In Lemma C.7 choosing $k^* = 0$ and $k^\dagger = \infty$ we obtain

$$\begin{aligned}\mathbf{B}^{(\ell)} &\preceq (\mathbf{I} - \gamma^{(\ell-1)}\mathbf{H})^K \cdot \mathbf{B}^{(\ell-1)} \cdot (\mathbf{I} - \gamma^{(\ell-1)}\mathbf{H})^K + 5\alpha\gamma^{(\ell-1)} \log K \cdot \big\langle \mathbf{H}, \mathbf{B}^{(\ell-1)} \big\rangle \cdot \mathbf{I} \\ &\preceq \mathbf{B}^{(\ell-1)} + 5\alpha\gamma^{(\ell-1)} \log K \cdot \big\langle \mathbf{H}, \mathbf{B}^{(\ell-1)} \big\rangle \cdot \mathbf{I},\end{aligned}$$

which implies that

$$\big\langle \mathbf{H}, \mathbf{B}^{(\ell)} \big\rangle \le \big(1 + 5\alpha\gamma^{(\ell-1)} \operatorname{tr}(\mathbf{H}) \log K\big) \cdot \big\langle \mathbf{H}, \mathbf{B}^{(\ell-1)} \big\rangle.$$

The above inequality provides us with a recursion about the bias iterates that would not blow up:

$$\begin{aligned}\big\langle \mathbf{H}, \mathbf{B}^{(L)} \big\rangle &\le \prod_{\ell=3}^{L} \big(1 + 5\alpha\gamma^{(\ell-1)} \operatorname{tr}(\mathbf{H}) \log K\big) \cdot \big\langle \mathbf{H}, \mathbf{B}^{(2)} \big\rangle \\ &\le e^{\sum_{\ell=3}^{L} 5\alpha\gamma^{(\ell-1)} \operatorname{tr}(\mathbf{H}) \log K} \cdot \big\langle \mathbf{H}, \mathbf{B}^{(2)} \big\rangle \\ &\le e \cdot \big\langle \mathbf{H}, \mathbf{B}^{(2)} \big\rangle,\end{aligned}$$

where the last inequality is because $\sum_{\ell=3}^{L} \gamma^{(\ell-1)} \le \gamma_0/2$ and $\alpha\gamma_0 \operatorname{tr}(\mathbf{H}) \log K < 1/3$. □

Lemma C.9. *Suppose Assumptions 3.1 and 3.2 hold. Consider (27) and (5). Suppose $\gamma_0 < 1/(3\alpha \operatorname{tr}(\mathbf{H}) \log(s+K))$. We have*

$$\begin{aligned}&\big\langle \mathbf{H}, \mathbf{B}^{(2)} \big\rangle \le 12 \cdot \Big\langle \frac{1}{\gamma_0 K}\mathbf{I}_{0:k^*} + \mathbf{H}_{k^*:\infty}, (\mathbf{I} - \gamma\mathbf{H})^{2(s+K)}\mathbf{B}_0 \Big\rangle + \\ &108\alpha \log(s+K) \cdot \Big\langle \frac{1}{\gamma_0(s+K)}\mathbf{I}_{0:k^\dagger} + \mathbf{H}_{k^\dagger:\infty}, \mathbf{B}_0 \Big\rangle \cdot \Big(\frac{k^*}{K} + \gamma_0 \sum_{k^* \le i < k^\dagger} \lambda_i + \gamma_0^2 (s+K) \sum_{i > k^\dagger} \lambda_i^2\Big),\end{aligned}$$

where k^* and $k^\dagger$ can be arbitrary.

Proof. According to Lemma C.6, we have

$$\big\langle \mathbf{H}, \mathbf{B}^{(2)} \big\rangle \le 6 \cdot \Big\langle \frac{1}{\gamma^{(1)} K}\mathbf{I}_{0:k^*} + \mathbf{H}_{k^*:\infty}, \mathbf{B}^{(1)} \Big\rangle \le 12 \cdot \Big\langle \frac{1}{\gamma_0 K}\mathbf{I}_{0:k^*} + \mathbf{H}_{k^*:\infty}, \mathbf{B}^{(1)} \Big\rangle.$$

On the other hand, in Lemma C.7 choosing $k^* = k^\dagger$, we have

$$\begin{aligned}&\mathbf{B}^{(1)} \le (\mathbf{I} - \gamma_0\mathbf{H})^{s+K} \cdot \mathbf{B}_0 \cdot (\mathbf{I} - \gamma_0\mathbf{H})^{s+K} + \\ &9\alpha\gamma_0^2 (s+K) \log(s+K) \cdot \Big\langle \frac{1}{\gamma_0(s+K)}\mathbf{I}_{0:k^\dagger} + \mathbf{H}_{k^\dagger:\infty}, \mathbf{B}_0 \Big\rangle \cdot \Big(\frac{1}{\gamma_0(s+K)}\mathbf{I}_{0:k^\dagger} + \mathbf{H}_{k^\dagger:\infty}\Big).\end{aligned}$$

Combining these two inequalities yields:

$$\begin{aligned}\big\langle \mathbf{H}, \mathbf{B}^{(2)} \big\rangle \le 12 \cdot \Big\langle \frac{1}{\gamma_0 K}\mathbf{I}_{0:k^*} + \mathbf{H}_{k^*:\infty}, (\mathbf{I} - \gamma_0\mathbf{H})^{2(s+K)}\mathbf{B}_0 \Big\rangle + \\ 108\alpha\gamma_0^2 (s+K) \log(s+K) \cdot \Big\langle \frac{1}{\gamma_0(s+K)}\mathbf{I}_{0:k^\dagger} + \mathbf{H}_{k^\dagger:\infty}, \mathbf{B}_0 \Big\rangle \times \\ \underbrace{\Big\langle \frac{1}{\gamma_0 K}\mathbf{I}_{0:k^*} + \mathbf{H}_{k^*:\infty}, \ \frac{1}{\gamma_0(s+K)}\mathbf{I}_{0:k^\dagger} + \mathbf{H}_{k^\dagger:\infty} \Big\rangle}_{(*)}.\end{aligned}$$

The proof is completed by noting that

$$(*) \le \frac{k^*}{\gamma_0^2 K(s+K)} + \frac{1}{\gamma_0(s+K)} \sum_{k^* < i \le k^\dagger} \lambda_i + \sum_{i > k^\dagger} \lambda_i^2.$$

where $k^\dagger \ge k^*$ and k^* and $k^\dagger$ are otherwise arbitrary. □

Theorem C.10 (A bias upper bound)**.** *Suppose Assumptions 3.1 and 3.2 hold. Consider (5). Suppose* $\gamma_0 < 1/(3\alpha\operatorname{tr}(\mathbf{H})\log(s+K))$*. We have*

$$\langle \mathbf{H}, \mathbf{B}_N \rangle \le 12e \cdot \Big\langle \frac{1}{\gamma_0 K}\mathbf{I}_{0:k^*} + \mathbf{H}_{k^*:\infty}, (\mathbf{I}-\gamma_0\mathbf{H})^{2(s+K)}\mathbf{B}_0 \Big\rangle + 108e\alpha\log(s+K)\cdot\Big\langle \frac{1}{\gamma_0(s+K)}\mathbf{I}_{0:k^\dagger} + \mathbf{H}_{k^\dagger:\infty}, \mathbf{B}_0\Big\rangle \cdot \Big(\frac{k^*}{K} + \gamma_0 \sum_{k^*\le i<k^\dagger}\lambda_i + \gamma_0^2(s+K)\sum_{i>k^\dagger}\lambda_i^2\Big),$$

where k^* *and* $k^\dagger$ *can be arbitrary.*

Proof. This is by Lemmas C.8 and C.9. □

C.4. Proof of Theorem 4.1

Proof of Theorem 4.1. This is by combining Lemma B.2, Theorems C.2 and C.10, and set $R^2 = \alpha\operatorname{tr}(\mathbf{H})$. □

C.5. Proof of Corollary 1

Proof of Corollary 1. For all these examples one can verify that $\operatorname{tr}(\mathbf{H}) \eqsim 1$. Therefore $\gamma_0 \eqsim 1/\log N$.

According to the optimal choice of k^* and $k^\dagger$, we can verify that

$$\frac{\big\|(\mathbf{I}-\gamma_0\mathbf{H})^{s+K}(\mathbf{w}_0-\mathbf{w}^*)\big\|^2_{\mathbf{I}_{0:k^*}}}{\gamma_0 K} + \big\|(\mathbf{I}-\gamma_0\mathbf{H})^{s+K}(\mathbf{w}_0-\mathbf{w}^*)\big\|^2_{\mathbf{H}_{k^*:\infty}} \lesssim \frac{\big\|(\mathbf{w}_0-\mathbf{w}^*)\big\|_2^2}{\gamma_0 K} \lesssim \frac{\log^2 N}{N},$$

and that

$$\log(s+K)\cdot\left(\frac{\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{I}_{0:k^\dagger}}}{\gamma_0(s+K)} + \|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{H}_{k^\dagger:\infty}}\right) \lesssim \log N\cdot\frac{\big\|(\mathbf{w}_0-\mathbf{w}^*)\big\|_2^2}{\gamma_0(s+K)} \lesssim \frac{\log^2 N}{N},$$

therefore in Theorem 4.1 we have

$$\begin{aligned}\texttt{ExcessRisk} &\le \texttt{BiasError} + \texttt{VarianceError}\\ &\lesssim \frac{\log^2 N}{N} + \frac{\log^2 N}{N}\cdot(*) + (*)\\ &\lesssim \max\Big\{\frac{\log^2 N}{N}, (*)\Big\},\end{aligned}$$

where

$$\begin{aligned}(*) &= \frac{k^*}{K} + \gamma_0\sum_{k^*<i\le k^\dagger}\lambda_i + \gamma_0^2(s+K)\sum_{i>k^\dagger}\lambda_i^2\\ &\eqsim \frac{k^*\log N}{N} + \frac{1}{\log N}\cdot\sum_{k^*<i\le k^\dagger}\lambda_i + \frac{N}{\log^2 N}\cdot\sum_{i>k^\dagger}\lambda_i^2.\end{aligned}$$

We next exam the order of $\log^2 N/N$ vs. $(*)$.

1. By definitions we have

$$k^* \eqsim \left(\frac{N}{\log^2 N}\right)^{\frac{1}{1+r}}, \quad k^\dagger \eqsim \left(\frac{N}{\log N}\right)^{\frac{1}{1+r}},$$

therefore we have

$$\begin{aligned}(*) &\eqsim \left(\frac{N}{\log^2 N}\right)^{\frac{1}{1+r}}\cdot\frac{\log N}{N} + \frac{1}{\log N}\cdot\left(\frac{N}{\log^2 N}\right)^{\frac{-r}{1+r}} + \frac{N}{\log^2 N}\cdot\left(\frac{N}{\log N}\right)^{\frac{-1-2r}{1+r}}\\ &\eqsim (\log N)^{\frac{r-1}{1+r}}\cdot N^{\frac{-r}{1+r}} + (\log N)^{\frac{-1}{1+r}}\cdot N^{\frac{-r}{1+r}} \eqsim (\log N)^{\frac{r-1}{1+r}}\cdot N^{\frac{-r}{1+r}}.\end{aligned}$$

This implies that $\texttt{ExcessRisk} \lesssim (\log N)^{\frac{r-1}{1+r}}\cdot N^{\frac{-r}{1+r}}$.

2. By definitions we have
$$k^* \eqsim N \cdot (\log N)^{-2-r}, \quad k^\dagger \eqsim N \cdot (\log N)^{-1-r},$$
therefore we have
$$\begin{aligned}(*) &\eqsim (\log N)^{-1-r} + \frac{1}{\log N} \cdot (\log k^*)^{1-r} + \frac{N}{\log^2 N} \cdot \left((k^\dagger)^{-1} \cdot (\log k^\dagger)^{-2r}\right) \\ &\eqsim (\log N)^{-1-r} + (\log N)^{-r} + (\log N)^{-1-r} \eqsim (\log N)^{-r}.\end{aligned}$$
This implies that $\texttt{ExcessRisk} \lesssim (\log N)^{-r}$.

3. By definitions we have
$$k^* \eqsim \log N, \quad k^\dagger \eqsim \log N,$$
therefore we have
$$(*) \eqsim \frac{\log^2 N}{N} + \frac{1}{\log N} \cdot 2^{-k^*} + \frac{N}{\log^2 N} \cdot 2^{-2k^\dagger} \eqsim \frac{\log^2 N}{N}.$$
This implies that $\texttt{ExcessRisk} \lesssim \log^2 N/N$.

□

D. Proof of Lower Bound

D.1. Variance Lower Bound

Theorem D.1 (A variance lower bound)**.** *Suppose Assumptions 3.1 and 3.3' hold. Consider (6). Let $K = (N-s)/\log(N-s)$. Suppose $s \geq 0$, $K \geq 10$ and $\gamma_0 < 1/\lambda_1$. We have*
$$\langle \mathbf{H}, \mathbf{C}_N \rangle \geq \frac{\sigma^2}{400} \left(\frac{k^*}{K} + \gamma_0 \sum_{k^* < i \leq k^\dagger} \lambda_i + \gamma_0^2 (s+K) \sum_{i > k^\dagger} \lambda_i^2 \right),$$
where $k^* := \max\{k : \lambda_k \geq 1/(\gamma_0 K)\}$ and $k^\dagger := \max\{k : \lambda_k \geq 1/(\gamma_0(s+K))\}$.

Proof. Notice that
$$\begin{aligned}\mathbf{C}_t &= (\mathcal{I} - \gamma_t \widetilde{\mathcal{T}}_t) \circ \mathbf{C}_{t-1} + \gamma_t^2 (\mathcal{M} - \widetilde{\mathcal{M}}) \circ \mathbf{C}_{t-1} + \gamma_t^2 \sigma^2 \mathbf{H} \\ &\succeq (\mathcal{I} - \gamma_t \widetilde{\mathcal{T}}_t) \circ \mathbf{C}_{t-1} + \gamma_t^2 \sigma^2 \mathbf{H}.\end{aligned}$$
Solving the recursion we obtain
$$\begin{aligned}\mathbf{C}_N &\succeq \sigma^2 \sum_{t=1}^N \gamma_t^2 \prod_{i=t+1}^N (\mathcal{I} - \gamma_i \widetilde{\mathcal{T}}_i) \circ \mathbf{H} \\ &= \sigma^2 \sum_{t=1}^N \gamma_t^2 \prod_{i=t+1}^N (\mathbf{I} - \gamma_i \mathbf{H})^2 \mathbf{H} \\ &\succeq \sigma^2 \underbrace{\sum_{t=1}^N \gamma_t^2 \prod_{i=t+1}^N (\mathbf{I} - 2\gamma_i \mathbf{H}) \mathbf{H}}_{(*)}.\end{aligned} \tag{28}$$
Now recalling (3), we have
$$(*) = \gamma_0^2 \sum_{i=1}^{s+K} \Big(\mathbf{I} - 2\gamma_0 \mathbf{H}\Big)^{s+K-i} \prod_{j=1}^{L-1} \Big(\mathbf{I} - \frac{\gamma_0}{2^{j-1}} \mathbf{H}\Big)^K \mathbf{H}$$

$$
\begin{aligned}
&+\sum_{\ell=1}^{L-1}\Big(\frac{\gamma_0}{2^\ell}\Big)^2\sum_{i=1}^{K}\Big(\mathbf{I}-\frac{\gamma_0}{2^{\ell-1}}\mathbf{H}\Big)^{K-i}\prod_{j=\ell+1}^{L-1}\Big(\mathbf{I}-\frac{\gamma_0}{2^{j-1}}\mathbf{H}\Big)^{K}\mathbf{H}\\
&=\frac{\gamma_0}{2}\left(\mathbf{I}-\Big(\mathbf{I}-2\gamma_0\mathbf{H}\Big)^{s+K}\right)\left(\prod_{j=1}^{L-1}\Big(\mathbf{I}-\frac{\gamma_0}{2^{j-1}}\mathbf{H}\Big)\right)^{K}\\
&\quad+\sum_{\ell=1}^{L-1}\frac{\gamma_0}{2^{\ell+1}}\left(\mathbf{I}-\Big(\mathbf{I}-\frac{\gamma_0}{2^{\ell-1}}\mathbf{H}\Big)^{K}\right)\left(\prod_{j=\ell+1}^{L-1}\Big(\mathbf{I}-\frac{\gamma_0}{2^{j-1}}\mathbf{H}\Big)\right)^{K}\\
&\geq\frac{\gamma_0}{2}\left(\mathbf{I}-\Big(\mathbf{I}-2\gamma_0\mathbf{H}\Big)^{s+K}\right)\Big(\mathbf{I}-2\gamma_0\mathbf{H}\Big)^{K}\\
&\quad+\sum_{\ell=1}^{L-1}\frac{\gamma_0}{2^{\ell+1}}\left(\mathbf{I}-\Big(\mathbf{I}-\frac{\gamma_0}{2^{\ell-1}}\mathbf{H}\Big)^{K}\right)\Big(\mathbf{I}-\frac{\gamma_0}{2^{\ell-1}}\mathbf{H}\Big)^{K},
\end{aligned}
\tag{29}
$$

where we understand $\prod_{j=L}^{L-1}(\cdot)=1$, and the last inequality is because for every $\ell\geq 0$,

$$
\prod_{j=\ell+1}^{L-1}\Big(\mathbf{I}-\frac{\gamma_0}{2^{j-1}}\mathbf{H}\Big)\geq\mathbf{I}-\sum_{j=\ell+1}^{L-1}\frac{\gamma_0}{2^{j-1}}\mathbf{H}\geq\mathbf{I}-\frac{\gamma_0}{2^{\ell-1}}\mathbf{H},
$$

where we understand $\sum_{j=L}^{L-1}(\cdot)=0$. Define a scalar function

$$
f(x):=\frac{x}{2}\cdot\Big(1-\big(1-2x\big)^{s+K}\Big)\cdot\big(1-2x\big)^{K}+\sum_{\ell=1}^{L-1}\frac{x}{2^{\ell+1}}\cdot\Big(1-\big(1-\frac{x}{2^{\ell-1}}\big)^{K}\Big)\cdot\big(1-\frac{x}{2^{\ell-1}}\big)^{K},
$$

and apply it to $\gamma_0\mathbf{H}$ entry-wisely, then according to Lemma D.2, we have

$$
f(\gamma_0\mathbf{H})\succeq\frac{1}{400K}\mathbf{I}_{0:k^*}+\frac{\gamma_0}{40}\mathbf{H}_{k^*:k^\dagger}+\frac{\gamma_0^2(s+K)}{40}\mathbf{H}^2_{k^\dagger:\infty},
$$

where $k^*:=\max\{k:\lambda_k\geq 1/(\gamma_0 K)\}$ and $k^\dagger:=\max\{k:\lambda_k\geq 1/(\gamma_0(s+K))\}$. Now using (28) and (29) we obtain

$$
\mathbf{C}_N\succeq\sigma^2\cdot f(\gamma_0\mathbf{H})\cdot\mathbf{H}^{-1}\succeq\frac{\sigma^2}{400}\left(\frac{1}{K}\mathbf{H}^{-1}_{0:k^*}+\gamma_0\mathbf{I}_{k^*:k^\dagger}+\gamma_0^2(s+K)\mathbf{H}_{k^\dagger:\infty}\right),
$$

and as a consequence,

$$
\langle\mathbf{H},\mathbf{C}_N\rangle\geq\frac{\sigma^2}{400}\left(\frac{k^*}{K}+\gamma_0\sum_{k^*<i\leq k^\dagger}\lambda_i+\gamma_0^2(s+K)\sum_{i>k^\dagger}\lambda_i^2\right),
$$

where $k^*:=\max\{k:\lambda_k\geq 1/(\gamma_0 K)\}$ and $k^\dagger:=\max\{k:\lambda_k\geq 1/(\gamma_0(s+K))\}$. □

Lemma D.2. *Suppose $s\geq 0$, $K\geq 10$ and $x\in(0,1]$. For the scalar function*

$$
f(x):=\frac{x}{2}\cdot\Big(1-\big(1-2x\big)^{s+K}\Big)\cdot\big(1-2x\big)^{K}+\sum_{\ell=1}^{L-1}\frac{x}{2^{\ell+1}}\cdot\Big(1-\big(1-\frac{x}{2^{\ell-1}}\big)^{K}\Big)\cdot\big(1-\frac{x}{2^{\ell-1}}\big)^{K},
$$

we have

$$
f(x)\geq\begin{cases}(s+K)x^2/40, & 0<x<1/(s+K);\\ x/40, & 1/(s+K)\leq x<1/K;\\ 1/(400K), & 1/K\leq x\leq 1.\end{cases}
$$

Proof. We prove each part of the lower bound separately.

- For $x \in (0, 1/(s+K))$ and $K \geq 10$, we have $(1-2x)^{s+K} \leq (1-x)^{s+K} \leq 1-(s+K)x/2$ and $(1-2x)^K \geq (1-2/(s+K))^K \geq (1-2/K)^K \geq (1-2/10)^{10} \geq \frac{1}{10}$, which yield

$$f(x) \geq \frac{x}{2} \cdot \frac{(s+K)x}{2} \cdot \frac{1}{10} = \frac{(s+K)x^2}{40}.$$

- For $x \in [1/(s+K), 1/K)$ and $K \geq 10$, we have $(1-2x)^{s+K} \leq (1-2/(s+K))^{s+K} \leq 1/e^2$ and $(1-2x)^K \geq (1-2/K)^K \geq (1-2/10)^{10} \geq \frac{1}{10}$, which yield

$$f(x) \geq \frac{x}{2} \cdot \big(1 - \frac{1}{e^2}\big) \cdot \frac{1}{10} \geq \frac{x}{40}.$$

- For $x \in [1/K, 1]$, there is an $\ell^* := \lfloor \log(Kx) \rfloor \in [0, L)$, such that $2^{\ell^*}/K \leq x < 2^{\ell^*+1}/K$, which yields

$$\begin{aligned}
f(x) &\geq \sum_{\ell=0}^{L-1} \frac{x}{2^{\ell+1}} \cdot \Big(1 - \big(1 - \frac{x}{2^{\ell-1}}\big)^K\Big) \cdot \big(1 - \frac{x}{2^{\ell-1}}\big)^K \qquad \text{(since } s+K \geq K\text{)} \\
&\geq \frac{x}{2^{\ell^*+1}} \cdot \Big(1 - \big(1 - \frac{x}{2^{\ell^*-1}}\big)^K\Big) \cdot \big(1 - \frac{x}{2^{\ell^*-1}}\big)^K \\
&\geq \frac{1}{2K} \cdot \Big(1 - \big(1 - \frac{2}{K}\big)^K\Big) \cdot \big(1 - \frac{4}{K}\big)^K \\
&\geq \frac{1}{2K} \cdot \Big(1 - \frac{1}{e^2}\Big) \cdot \Big(1 - \frac{4}{10}\Big)^{10} \geq \frac{1}{400K}. \qquad \text{(since } K \geq 10\text{)}
\end{aligned}$$

□

D.2. Bias Lower Bound

We now build a lower bound for the bias error.

Theorem D.3. *Suppose Assumptions 3.1, 3.2 and 3.3' hold. Consider (5). Let $K = (N-s)/\log(N-s)$. Suppose $s \geq 0$, $K \geq 10$ and $\gamma_0 < 1/\lambda_1$. We have*

$$\begin{aligned}
\langle \mathbf{H}, \mathbf{B}_N \rangle \geq \big\langle \mathbf{H}, (\mathbf{I} - \gamma_0 \mathbf{H})^{2(s+2K)} \mathbf{B}_0 \big\rangle \\
+ \frac{\beta}{1200} \cdot \langle \mathbf{H}_{k^\dagger:\infty}, \mathbf{B}_0 \rangle \cdot \left(\frac{k^*}{K} + \gamma_0 \sum_{k^* < i \leq k^\dagger} + \gamma_0^2 (s+K) \sum_{i > k^\dagger} \lambda_i^2 \right),
\end{aligned}$$

where $k^* := \max\{k : \lambda_k \geq 1/(\gamma_0 K)\}$ and $k^\dagger := \max\{k : \lambda_k \geq 1/(\gamma_0 (s+K))\}$.

Proof. Starting from (5), we have

$$\begin{aligned}
\mathbf{B}_n &= (\mathcal{I} - \gamma_n \widetilde{\mathcal{T}}_n) \circ \mathbf{B}_{n-1} + \gamma_n^2 (\mathcal{M} - \widetilde{\mathcal{M}}) \circ \mathbf{B}_{n-1} \\
&\succeq (\mathcal{I} - \gamma_n \widetilde{\mathcal{T}}_n) \circ \mathbf{B}_{n-1} + \beta \gamma_n^2 \cdot \mathbf{H} \cdot \langle \mathbf{H}, \mathbf{B}_{n-1} \rangle \\
&\succeq (\mathcal{I} - \gamma_n \widetilde{\mathcal{T}}_n) \circ \mathbf{B}_{n-1},
\end{aligned} \tag{30}$$

recursively solving this, we obtain a crude lower bound on $\mathbf{B}_n$:

$$\mathbf{B}_n \succeq \prod_{t=1}^{n} (\mathcal{I} - \gamma_t \widetilde{\mathcal{T}}_t) \circ \mathbf{B}_0 \succeq \prod_{t=1}^{N} (\mathcal{I} - \gamma_t \widetilde{\mathcal{T}}_t) \circ \mathbf{B}_0, \quad \text{for } n = 1, \ldots, N.$$

This gives us a crude lower bound on $\langle \mathbf{H}, \mathbf{B}_n \rangle$ for $n = 1, \ldots, N$:

$$\langle \mathbf{H}, \mathbf{B}_n \rangle \geq \Big\langle \prod_{t=1}^{N} (\mathcal{I} - \gamma_t \widetilde{\mathcal{T}}_t) \circ \mathbf{H},\ \mathbf{B}_0 \Big\rangle = \Big\langle \prod_{t=1}^{N} (\mathbf{I} - \gamma_t \mathbf{H})^2 \mathbf{H},\ \mathbf{B}_0 \Big\rangle.$$

Bring this into (30), we have

$$\mathbf{B}_n \succeq (\mathcal{I} - \gamma_n \widetilde{\mathcal{T}}_n) \circ \mathbf{B}_{n-1} + \beta\gamma_n^2 \cdot \mathbf{H} \cdot \Big\langle \prod_{t=1}^{N} (\mathbf{I} - \gamma_t \mathbf{H})^2 \mathbf{H},\ \mathbf{B}_0 \Big\rangle, \quad \text{for } n = 1, \dots, N,$$

recursively solving which yields:

$$\mathbf{B}_N \succeq \prod_{t=1}^{n} (\mathcal{I} - \gamma_t \widetilde{\mathcal{T}}_t) \circ \mathbf{B}_0 + \beta \Big\langle \prod_{t=1}^{N} (\mathbf{I} - \gamma_t \mathbf{H})^2 \mathbf{H},\ \mathbf{B}_0 \Big\rangle \cdot \underbrace{\sum_{t=1}^{N} \gamma_t^2 \prod_{i=t+1}^{N} (\mathcal{I} - \gamma_i \widetilde{\mathcal{T}}_i) \circ \mathbf{H}}_{(*)}. \tag{31}$$

Noting that here the term $(*)$ in (31) is exactly the term $(*)$ appeared in (28) in Theorem D.1, therefore by repeating the analysis in Theorem D.1 we know that

$$(*) \succeq \frac{1}{400} \left(\frac{1}{K} \mathbf{H}_{0:k^*}^{-1} + \gamma_0 \mathbf{I}_{k^*:k^\dagger} + \gamma_0^2 (s+K) \mathbf{H}_{k^\dagger:\infty} \right),$$

where $k^* := \max\{k : \lambda_k \geq 1/(\gamma_0 K)\}$ and $k^\dagger := \max\{k : \lambda_k \geq 1/(\gamma_0(s+K))\}$. As a consequence we have

$$\langle \mathbf{H}, (*) \rangle \geq \frac{1}{400} \left(\frac{k^*}{K} + \gamma_0 \sum_{k^* < i \leq k^\dagger} + \gamma_0^2 (s+K) \sum_{i > k^\dagger} \lambda_i^2 \right). \tag{32}$$

Back to (31), taking inner product with $\mathbf{H}$ yields

$$\begin{aligned}
\langle \mathbf{H}, \mathbf{B}_N \rangle &\geq \Big\langle \prod_{t=1}^{N} (\mathcal{I} - \gamma_t \widetilde{\mathcal{T}}_t) \circ \mathbf{H},\ \mathbf{B}_0 \Big\rangle + \beta \cdot \Big\langle \prod_{t=1}^{N} (\mathbf{I} - \gamma_t \mathbf{H})^2 \mathbf{H},\ \mathbf{B}_0 \Big\rangle \cdot \langle \mathbf{H}, (*) \rangle \\
&= \underbrace{\Big\langle \prod_{t=1}^{N} (\mathbf{I} - \gamma_t \mathbf{H})^2 \mathbf{H},\ \mathbf{B}_0 \Big\rangle}_{(**)} + \beta \cdot \underbrace{\Big\langle \prod_{t=1}^{N} (\mathbf{I} - \gamma_t \mathbf{H})^2 \mathbf{H},\ \mathbf{B}_0 \Big\rangle}_{(**)} \cdot \langle \mathbf{H}, (*) \rangle .
\end{aligned} \tag{33}$$

We next bound $(**)$. Recall (3), we have

$$\begin{aligned}
(**) &= (\mathbf{I} - \gamma_0 \mathbf{H})^{2(s+K)} \cdot \prod_{\ell=1}^{L-1} \big(\mathbf{I} - \frac{\gamma_0}{2^\ell} \mathbf{H}\big)^{2K} \cdot \mathbf{H} \\
&\succeq (\mathbf{I} - \gamma_0 \mathbf{H})^{2(s+K)} \cdot \big(\mathbf{I} - \sum_{\ell=1}^{L-1} \frac{\gamma_0}{2^\ell} \mathbf{H}\big)^{2K} \cdot \mathbf{H} \\
&\succeq (\mathbf{I} - \gamma_0 \mathbf{H})^{2(s+K)} \cdot \big(\mathbf{I} - \gamma_0 \mathbf{H}\big)^{2K} \cdot \mathbf{H} \\
&= (\mathbf{I} - \gamma_0 \mathbf{H})^{2(s+2K)} \mathbf{H},
\end{aligned} \tag{34}$$

Noticing that for $K \geq 10$ and $x \in (0, 1/(s+K))$,

$$(1-x)^{2(s+2K)} \geq (1-x)^{s+K} \geq \Big(1 - \frac{1}{s+K}\Big)^{s+K} \geq \big(1 - \frac{1}{10}\big)^{10} \geq \frac{1}{3},$$

we can further lower bound $(**)$ with $k^\dagger := \max\{k : \lambda_k \geq 1/(\gamma_0(s+K))\}$:

$$(**) \succeq (\mathbf{I} - \gamma_0 \mathbf{H})^{2(s+2K)} \mathbf{H} \succeq \frac{1}{3} \mathbf{H}_{k^\dagger:\infty} \tag{35}$$

Bringing (32), (34) and (35) into (33) completes the proof:

$$\begin{aligned}
\langle \mathbf{H}, \mathbf{B}_N \rangle \geq\ & \big\langle \mathbf{H}, (\mathbf{I} - \gamma_0 \mathbf{H})^{2(s+2K)} \mathbf{B}_0 \big\rangle \\
& + \frac{\beta}{1200} \cdot \langle \mathbf{H}_{k^\dagger:\infty}, \mathbf{B}_0 \rangle \cdot \left(\frac{k^*}{K} + \gamma_0 \sum_{k^* < i \leq k^\dagger} + \gamma_0^2 (s+K) \sum_{i > k^\dagger} \lambda_i^2 \right),
\end{aligned}$$

where $k^* := \max\{k : \lambda_k \geq 1/(\gamma_0 K)\}$ and $k^\dagger := \max\{k : \lambda_k \geq 1/(\gamma_0(s+K))\}$. □

D.3. Proof of Theorem 4.2

Proof of Theorem 4.2. This is by combining Lemma B.3, Theorems D.1 and D.3. □

E. Proof for Polynomially Decaying Stepsize

Recall that the polynomially decaying stepsize satisfies the following rule:

$$\gamma_t = \begin{cases} \gamma_0, & 1 \le t \le s; \\ \gamma_0/(t-s)^a, & s < t \le N. \end{cases} \tag{36}$$

E.1. Proof of the Lower Bound of Variance Error

Lemma E.1. *Suppose $\gamma_0 < 1/(4\lambda_1)$ and apply polynomially decaying stepsize, then it holds that*

Case 1: *$0 \le a < 1$. Let $k^* = \max\{k : \gamma_0\lambda_k \ge (1-a)/[2(N-s)^{a-1}]\}$ and $k^\dagger = \max\{k : \gamma_0\lambda_k \ge 1/(2s)\}$, we have*

$$\langle \mathbf{H}, \mathbf{C}_N \rangle \ge \sigma^2 \cdot \bigg(\sum_{i \le k^*} \frac{(1-a)\cdot\gamma_0\lambda_i}{N^a} \vee \frac{(1-a)^2 a \log(N)}{16eN} + \sum_{k^*+1 \le i \le k^\dagger} \frac{\gamma_0\lambda_i}{4e^2} + \sum_{i \ge k^\dagger+1} \frac{s\gamma^2\lambda_i^2}{2e^2} \bigg).$$

Case 2: *$a = 1$. Let $k^* = \max\{k : \gamma_0\lambda_k \ge 1/(2+2\log(N-s-1))\}$ and $k^\dagger = \max\{k : \gamma_0\lambda_k \ge 1/(2s)\}$, we have*

$$\langle \mathbf{H}, \mathbf{C}_N \rangle \ge \sigma^2 \cdot \bigg(\sum_{i \le k^*} \frac{\gamma^2\lambda_i^2}{N^{4\gamma_0\lambda_i}} + \sum_{k^*+1 \le i \le k^\dagger} \frac{\gamma_0\lambda_i}{4e^2} + \sum_{i \ge k^\dagger+1} \frac{s\gamma^2\lambda_i^2}{2e^2} \bigg).$$

Proof. Consider $\mathbf{C}_t$ defined in (6), we have

$$\begin{aligned} \mathbf{C}_N &= (\mathcal{I} - \gamma_N\widetilde{\mathcal{T}}) \circ \mathbf{C}_{N-1} + \gamma_N^2(\mathcal{M} - \widetilde{\mathcal{M}}) \circ \mathbf{C}_{N-1} + \gamma_N^2\sigma^2\mathbf{H} \\ &\succeq (\mathcal{I} - \gamma_N\widetilde{\mathcal{T}}) \circ \mathbf{C}_{N-1} + \gamma_N^2\sigma^2\mathbf{H} \\ &\succeq \sigma^2 \sum_{t=1}^N \gamma_t^2 \prod_{i=t+1}^N (\mathcal{I} - \gamma_i\widetilde{\mathcal{T}}) \circ \mathbf{H}, \end{aligned} \tag{37}$$

where in the last inequality we use the fact that $\mathbf{C}_0 = \mathbf{0}$. Then using the fact that $\mathbf{H}$ is a PSD matrix, it holds that

$$\begin{aligned} \langle \mathbf{H}, \mathbf{C}_N \rangle &\ge \sigma^2 \cdot \sum_{t=1}^N \gamma_t^2 \cdot \bigg\langle \prod_{i=t+1}^N (\mathcal{I} - \gamma_i\widetilde{\mathcal{T}}) \circ \mathbf{H}, \mathbf{H} \bigg\rangle \\ &= \sigma^2 \cdot \sum_{t=1}^N \gamma_t^2 \cdot \bigg\langle \prod_{i=t+1}^N (\mathbf{I} - \gamma_i\mathbf{H})^2\mathbf{H}, \mathbf{H} \bigg\rangle \\ &= \sigma^2 \sum_j \sum_{t=1}^N \gamma_t^2 \cdot \prod_{i=t+1}^N (1-\gamma_i\lambda_j)^2\lambda_j^2, \end{aligned}$$

where the second equality follows from the definition of $\widetilde{\mathcal{T}}$ and the fact that $\widetilde{\mathcal{T}} \circ \mathbf{A}$ is commute to $\mathbf{H}$ for any $\mathbf{A}$ that is commute to $\mathbf{H}$. Then it suffices to consider the following scalar function:

$$\begin{aligned} f(x) :&= \sum_{t=1}^N \gamma_t^2 \prod_{i=t+1}^N (1-\gamma_i x)^2 x^2 \\ &= \underbrace{\gamma^2 \sum_{t=1}^s \prod_{i=t+1}^N (1-\gamma_i x)^2 x^2}_{:= f_1(x)} + \underbrace{\sum_{t=s+1}^N \gamma_t^2 \prod_{i=t+1}^N (1-\gamma_i x)^2 x^2}_{:= f_2(x)} \end{aligned} \tag{38}$$

where we explicitly decompose the function $f(x)$ into the summation of two functions $f_1(x)$ and $f_2(x)$ according to the length of iterations that use stepsize γ.

Lower bound of $f_1(x)$. We first provide a lower bound of $f_1(x)$. Note that $f_1(x)$ can be rewritten as

$$\begin{aligned} f_1(x) &= \gamma^2 \cdot \sum_{t=1}^{s} (1-\gamma x)^{2(s-t)} \cdot \prod_{i=s+1}^{N} (1-\gamma_i x)^2 x^2 \\ &= \gamma^2 \cdot \prod_{i=s+1}^{N} (1-\gamma_i x)^2 x^2 \cdot \sum_{t=0}^{s-1} (1-\gamma x)^{2t}. \end{aligned}$$

Then note that for any $i \geq s+1$, we have $\gamma_i = \gamma/(i-s)^a$. Applying the fact that $(1-\gamma x)^2 \geq (1-2\gamma x)$ for any $\gamma x \leq 1/2$, it follows that

$$\begin{aligned} f_1(x) &\geq \gamma^2 \cdot \prod_{i=1}^{N-s} \left(1 - \frac{\gamma x}{i^a}\right)^2 x^2 \cdot \sum_{t=0}^{s-1} (1-2\gamma x)^t \\ &= \frac{\gamma x}{2} \cdot \big[1 - (1-2\gamma x)^s\big] \cdot \prod_{i=1}^{N-s} \left(1 - \frac{\gamma x}{i^a}\right)^2. \end{aligned}$$

Moreover, note that $\gamma x \leq 1/2$ implies that $(1-\gamma x/i^a) \geq e^{-2\gamma x/i^a}$ for any $i \geq 1$, we further have

$$f_1(x) \geq \frac{\gamma x}{2} \cdot \big[1 - (1-2\gamma x)^s\big] \cdot e^{-4\gamma x \cdot \sum_{i=1}^{N-s} i^{-a}}. \tag{39}$$

Note that

$$\sum_{i=1}^{N-s} i^{-a} = 1 + \sum_{i=2}^{N-s} i^{-a} \leq 1 + \int_{1}^{N-s-1} z^{-a} \mathrm{d}z = \begin{cases} 1 + \frac{(N-s-1)^{1-a}-1}{1-a}, & 0 \leq a < 1; \\ 1 + \log(N-s-1), & a = 1. \end{cases} \tag{40}$$

Therefore, plugging (40) into (39), it holds that

$$f_1(x) \geq \begin{cases} \frac{\gamma x}{2} \cdot \big[1 - (1-2\gamma x)^s\big] \cdot e^{-4\gamma x \cdot (N-s)^{1-a}/(1-a)}, & 0 \leq a < 1 \\ \frac{\gamma x}{2} \cdot \big[1 - (1-2\gamma x)^s\big] \cdot e^{-4\gamma x \cdot [1+\log(N-s)]}, & a = 1. \end{cases}$$

- **Case of $0 \leq a < 1$.** For the case of $0 \leq a < 1$, assume $s = \Omega((N-s)^{1-a})$, let $k^* = \max\{k : \gamma_0 \lambda_k \geq (1-a)/[2(N-s)^{1-a}]\}$ and $k^\dagger = \max\{k : \gamma_0 \lambda_k \geq 1/(2s)\}$, where it can be verified that $k^* \leq k^\dagger$). Then note that for any $k \leq k^\dagger$, we have

  $$1 - (1-2\gamma x)^s \geq 1 - (1-1/s)^s \geq 1/2$$

  and for any $k \geq k^\dagger + 1$,

  $$1 - (1-2\gamma x)^s \geq 1 - e^{-2s\gamma x} \geq 1 - (1 - s\gamma x) \geq s\gamma x$$

  where the second inequality holds since $e^{-x} \leq 1 - x/2$ for any $x \in [0,1]$. Besides, we also have for any $k \geq k^* + 1$,

  $$e^{-4\gamma x \cdot (N-s)^{1-a}/(1-a)} \geq e^{-2}.$$

  Besides, note that the $g(x) = xe^{-cx}$ first increases and then decreases as x increases, then for any for any $x \in \big[(1-a)/[2(N-s)^{1-a}], a(1-a)\log(N)/[4(N-s)^{1-a}]\big]$, we have

  $$\begin{aligned} f_1(x) &\geq \min\big\{f_1\big((1-a)/[2(N-s)^{1-a}]\big), f_1\big(a(1-a)\log(N)/[4(N-s)^{1-a}]\big)\big\} \\ &= \min\left\{\frac{1-a}{4e^2(N-s)^{1-a}}, \frac{a(1-a)\log(N)}{8(N-s)^{1-a}} \cdot e^{-a\log(N)}\right\} \end{aligned}$$

$$= \frac{a(1-a)\log(N)}{8N}.$$

Therefore, we can further define $k' = \max\{k : \gamma_0\lambda k \geq a(1-a)\log(N)/[4(N-s)^{a-1}]\}$ such that for any $k' < k \leq k^*$, it holds that

$$f_1(\lambda_k) \geq \frac{\gamma_0\lambda_k}{4} \cdot e^{-a\log(N)} = \frac{a(1-a)\log(N)}{8N}$$

Combining these bounds we can obtain that

$$\sum_i f_1(\lambda_i) \geq \sum_{i>k'} f_1(\lambda_i) \geq \sum_{k'<i\leq k^*} \frac{a(1-a)\log(N)}{8N} + \sum_{k^*<i\leq k^\dagger} \frac{\gamma_0\lambda_i}{4e^2} + \sum_{i\geq k^\dagger+1} \frac{s\gamma^2\lambda_i^2}{2e^2}. \tag{41}$$

- **Case of $a = 1$.** Similarly, when $a = 1$, we can redefine k^* as $k^* = \max\{k : \gamma_0\lambda_k \geq 1/(2 + 2\log(N - s - 1))\}$ and then similarly, it holds that

$$\sum_i f_1(\lambda_i) \geq \sum_{k\geq k^*+1} f_1(\lambda_i) \geq \sum_{k^*<i\leq k^\dagger} \frac{\gamma_0\lambda_i}{4e^2} + \sum_{i\geq k^\dagger+1} \frac{s\gamma^2\lambda_i^2}{2e^2}. \tag{42}$$

Lower bound of $f_2(x)$. Plugging the formula of the polynomially decaying stepsize, we have

$$\begin{aligned} f_2(x) &= \sum_{t=s+1}^{N} \gamma_t^2 \prod_{i=t+1}^{N} (1-\gamma_i x)^2 x^2 \\ &= \sum_{t=s+1}^{N} \frac{\gamma^2}{(t-s)^{2a}} \cdot \prod_{i=t+1}^{N} \left(1 - \frac{\gamma x}{(i-s)^a}\right)^2 x^2 \\ &= \sum_{t=1}^{N-s} \frac{\gamma^2}{t^{2a}} \cdot \prod_{i=t+1}^{N-s} \left(1 - \frac{\gamma x}{i^a}\right)^2 x^2 \end{aligned}$$

Similarly, for any $\gamma x \leq 1/2$, we have

$$\left(1 - \frac{\gamma x}{i^a}\right)^2 \geq 1 - \frac{2\gamma x}{i^a} \geq e^{-4\gamma x/i^a}.$$

Then it follows that

$$f_2(x) \geq \sum_{t=1}^{N-s} \frac{\gamma^2 x^2}{t^{2a}} \cdot e^{-4\gamma x \cdot \sum_{i=t+1}^{N-s} i^{-a}}. \tag{43}$$

- **Case of $0 \leq a < 1$**

We first consider the case of $0 \leq a < 1$, where two cases will be studied separately: (1) $\gamma x \leq (1-a)/[2(N-s)^{a-1}]$ and (2) $\gamma x > (1-a)/[2(N-s)^{a-1}]$. For the first case, it is clear that

$$4\gamma x \cdot \sum_{i=t+1}^{N-s} i^{-a} \leq 4\gamma x \cdot \sum_{i=1}^{N-s} i^{-a} \leq \frac{2(1-a)}{(N-s)^{1-a}} \cdot \frac{(N-s)^a}{(1-a)} = 2,$$

which implies that

$$f_2(x) \geq e^{-2} \cdot \gamma^2 x^2 \cdot \sum_{t=1}^{N-s} t^{-2a} \geq \frac{[1 + (N-s)^{1-2a}]\gamma^2 x^2}{2e^2}. \tag{44}$$

Then we can move to the case

$$4\gamma x \geq \frac{2(1-a)}{(N-s)^{1-a}}. \tag{45}$$

Let t^* be the index satisfying

$$\sum_{i=t^*+1}^{N-s} i^{-a} \leq \frac{1}{4\gamma x} \leq \sum_{i=t^*}^{N-s} i^{-a}. \tag{46}$$

Note that we have

$$\sum_{i=t^*+1}^{N-s} i^{-a} \geq \int_{t^*+1}^{N-s} z^{-a}\mathrm{d}z = \frac{(N-s)^{1-a} - (t^*+1)^{1-a}}{1-a}$$
$$\sum_{i=t^*}^{N-s} i^{-a} \leq \int_{t^*-1}^{N-s} z^{-a}\mathrm{d}z = \frac{(N-s)^{1-a} - (t^*-1)^{1-a}}{1-a}.$$

Plugging the above inequality into (46) gives

$$\frac{(N-s)^{1-a} - (t^*+1)^{1-a}}{1-a} \leq \frac{1}{4\gamma x} \leq \frac{(N-s)^{1-a} - (t^*-1)^{1-a}}{1-a},$$

which implies that

$$t^* \in \left[\left((N-s)^{1-a} - \frac{1-a}{4\gamma x}\right)^{\frac{1}{1-a}} - 1, \left[(N-s)^{1-a} - \frac{1-a}{4\gamma x}\right]^{\frac{1}{1-a}} + 1\right].$$

Note that

$$\begin{aligned}\left[(N-s)^{1-a} - \frac{1-a}{4\gamma x}\right]^{\frac{1}{1-a}} &= (N-s)\cdot\left[1 - \frac{1-a}{4\gamma x(N-s)^{1-a}}\right]^{\frac{1}{1-a}} \\ &\leq (N-s) - \frac{(1-a)\cdot(N-s)^a}{4\gamma x},\end{aligned}$$

where the inequality follows from (45) and the fact that $1/(1-a) \geq 1$. Therefore, it holds that

$$t^* \leq (N-s) - \frac{(1-a)\cdot(N-s)^a}{4\gamma x} + 1. \tag{47}$$

Therefore, applying the above inequality to (43) gives

$$\begin{aligned} f_2(x) &\geq \sum_{t=1}^{N-s} \frac{\gamma^2 x^2}{t^{2a}} \cdot e^{-4\gamma x\cdot\sum_{i=t+1}^{N-s} i^{-a}}. \\ &\geq \sum_{t=t^*}^{N-s} \frac{\gamma^2 x^2}{t^{2a}} \cdot e^{-4\gamma x\cdot\sum_{i=t+1}^{N} i^{-a}} \\ &\overset{(i)}{\geq} \sum_{t=t^*}^{N-s} \frac{\gamma^2 x^2}{N^{2a}} \cdot e^{-4\gamma x\cdot\sum_{i=t^*+1}^{N-s} i^{-a}} \\ &\overset{(ii)}{\geq} (N-s-t^*+1)\cdot\frac{\gamma^2 x^2}{(N-s)^{2a}}\cdot e^{-1} \\ &\overset{(iii)}{\geq} \frac{(1-a)\cdot\gamma x}{4e\cdot N^a}, \end{aligned} \tag{48}$$

where the (i) holds since $t \in [t^*, N-s]$, (ii) follows from the fact that $\sum_{i=t^*+1}^{N-s} i^{-a} \leq 1$, and (iii) follows from (47). Then combining (44) and 48 and set $k^* := \max\{k : \gamma_0\lambda_k \geq (1-a)/(2(N-s)^{1-a})$, we can get

$$\sum_i f_2(\lambda_i) \geq \sum_{i\leq k^*} \frac{(1-a)\cdot\gamma_0\lambda_i}{4e\cdot N^a} + \sum_{i\geq k^*+1} \frac{[1+(N-s)^{1-2a}]\gamma^2\lambda_i^2}{2e^2} \tag{49}$$

- **Case of $a = 1$.** Then considering the case of $a = 1$, where it holds that

$$\sum_{i=t+1}^{N-s} i^{-a} \leq \int_{t}^{N-s} z^{-a} \mathrm{d}z = \log(N-s) - \log(t).$$

Then the following holds according to (43),

$$\begin{aligned} f_2(x) &\geq \sum_{t=1}^{N-s} \frac{\gamma^2 x^2}{t^2} \cdot e^{-4\gamma x \cdot \sum_{i=t+1}^{N-s} i^{-a}} \\ &\geq \sum_{t=1}^{N-s} \frac{\gamma^2 x^2}{t^2} \cdot e^{-4\gamma x \cdot \left[\log(N-s) - \log(t)\right]} \\ &\geq \sum_{t=1}^{N-s} \frac{\gamma^2 x^2}{t^2} \cdot \left(\frac{t}{N-s}\right)^{4\gamma x}. \end{aligned}$$

Then note that for any $4\gamma x < 1$,

$$\sum_{t=1}^{N-s} \frac{\gamma^2 x^2}{t^2} \cdot \left(\frac{t}{N-s}\right)^{4\gamma x} = \frac{\gamma^2 x^2}{(N-s)^{4\gamma x}} \cdot \sum_{t=1}^{N-s} \frac{1}{t^{2-4\gamma x}} \geq \frac{\gamma^2 x^2}{(N-s)^{4\gamma x}},$$

which implies that

$$\sum_i f_2(\lambda_i) \geq \sum_i \frac{\gamma_0^2 \lambda_i^2}{(N-s)^{4\gamma_0 \lambda_i}}. \tag{50}$$

Now we can combine the derived lower bounds for $f_1(x)$ and $f_2(x)$ in (41), (42) (49), and (50), and obtain

- **Case of $0 \leq a < 1$.** Let $k^* = \max\{k : \gamma_0 \lambda_k \geq (1-a)/[2(N-s)^{a-1}]\}$ and $k^\dagger = \max\{k : \gamma_0 \lambda_k \geq 1/(2s)\}$, we have

$$\begin{aligned} \langle \mathbf{H}, \mathbf{C}_N \rangle &\geq \sigma^2 \cdot \sum_i [f_1(\lambda_i) + f_2(\lambda_i)] \\ &\geq \sigma^2 \cdot \bigg(\sum_{i \leq k^*} \frac{(1-a) \cdot \gamma_0 \lambda_i}{N^a} \vee \frac{(1-a)^2 a \log(N)}{16eN} + \sum_{k^*+1 \leq i \leq k^\dagger} \frac{\gamma_0 \lambda_i}{4e^2} + \sum_{i \geq k^\dagger + 1} \frac{s\gamma_0^2 \lambda_i^2}{2e^2} \bigg), \end{aligned}$$

where we use the fact that

$$\frac{(1-a)\gamma_0 \lambda_i}{N^a} \geq \frac{(1-a)^2 a \log(N)}{16eN}$$

for all $i \leq k'$ (please refer to (41) for the definition of k').

- **Case of $a = 1$.** Let $k^* = \max\{k : \gamma_0 \lambda_k \geq 1/(2 + 2\log(N-s-1))\}$ and $k^\dagger = \max\{k : \gamma_0 \lambda_k \geq 1/(2s)\}$, we have

$$\begin{aligned} \langle \mathbf{H}, \mathbf{C}_N \rangle &\geq \sigma^2 \cdot \sum_i [f_1(\lambda_i) + f_2(\lambda_i)] \\ &\geq \sigma^2 \cdot \bigg(\sum_{i \leq k^*} \frac{\gamma_0^2 \lambda_i^2}{N^{4\gamma_0 \lambda_i}} + \sum_{k^*+1 \leq i \leq k^\dagger} \frac{\gamma_0 \lambda_i}{4e^2} + \sum_{i \geq k^\dagger + 1} \frac{s\gamma_0^2 \lambda_i^2}{2e^2} \bigg). \end{aligned}$$

□

E.2. Proof of the Lower Bound of Bias Error

Lemma E.2. *If applying polynomially decaying stepsize with $\gamma < 1/(4\lambda_1)$, then it holds that*

- ***Case 1:*** $0 \le a < 1$. *Let* $k^* = \max\{k : \gamma_0\lambda_k \ge (1-a)/[2(N-s)^{a-1}]\}$ *and* $k^\dagger = \max\{k : \gamma_0\lambda_k \ge 1/(2s)\}$, *we have*

$$\langle \mathbf{H}, \mathbf{B}_N \rangle \ge \big\|(\mathbf{I}-\gamma\mathbf{H})^{s+2N^{1-a}/(1-a)} \cdot (\mathbf{w}_0 - \mathbf{w}^*)\big\|_{\mathbf{H}}^2 + e^{-4}\beta\langle \mathbf{H}_{k^\dagger:\infty}, \mathbf{B}_0\rangle \cdot \bigg(\sum_{i\le k^*} \frac{(1-a)\cdot\gamma_0\lambda_i}{N^a} + \sum_{k^*+1\le i\le k^\dagger} \frac{\gamma_0\lambda_i}{4e^2} + \sum_{i\ge k^\dagger+1} \frac{s\gamma_0^2\lambda_i^2}{2e^2}\bigg).$$

- ***Case 2:*** $a = 1$. *Let* $k^* = \max\{k : \gamma_0\lambda_k \ge 1/(2+2\log(N-s-1))\}$ *and* $k^\dagger = \max\{k : \gamma_0\lambda_k \ge 1/(2s)\}$, *we have*

$$\langle \mathbf{H}, \mathbf{B}_N \rangle \ge \big\|(\mathbf{I}-\gamma\mathbf{H})^{s+2\log(N)} \cdot (\mathbf{w}_0 - \mathbf{w}^*)\big\|_{\mathbf{H}}^2 + e^{-4}\beta\langle \mathbf{H}_{k^\dagger:\infty}, \mathbf{B}_0\rangle \cdot \bigg(\sum_{i\le k^*} \frac{\gamma_0^2\lambda_i^2}{N^{4\gamma_0\lambda_i}} + \sum_{k^*+1\le i\le k^\dagger} \frac{\gamma_0\lambda_i}{4e^2} + \sum_{i\ge k^\dagger+1} \frac{s\gamma_0^2\lambda_i^2}{2e^2}\bigg).$$

Proof. The proof of Lemma E.2 follows a similar idea of the proof of Thoerem D.3. In particular, by (31), we have

$$\mathbf{B}_N \succeq \prod_{t=1}^n (\mathcal{I}-\gamma_t\widetilde{\mathcal{T}}_t)\circ\mathbf{B}_0 + \beta\Big\langle \prod_{t=1}^N (\mathbf{I}-\gamma_t\mathbf{H})^2\mathbf{H},\ \mathbf{B}_0\Big\rangle \cdot \sum_{t=1}^N \gamma_t^2 \prod_{i=t+1}^N (\mathcal{I}-\gamma_i\widetilde{\mathcal{T}}_i)\circ\mathbf{H}. \tag{51}$$

Then we focus on the scalar function $g(x) = \prod_{t=1}^N (1-\gamma_t x)^2 x$ for all $x \le 1/(2\gamma)$. Specifically, using the inequality $(1-\gamma_t x)^2 \ge e^{-4\gamma_t x}$, we have

$$g(x) \ge e^{-4x\cdot\sum_{t=1}^N \gamma_t} \ge e^{-8s\gamma x},$$

where we use the assumption that $s\gamma \ge \sum_{t=s+1}^s \gamma_t$. Then let $k^\dagger := \max\{k : \gamma_0\lambda_k \ge 1/(2s)\}$, we have

$$\prod_{t=1}^N (\mathbf{I}-\gamma\mathbf{H})^2\mathbf{H} \succeq e^{-4}\cdot\mathbf{H}_{k^\dagger:\infty}.$$

Plugging the above inequality into (51) and multiplying by $\mathbf{H}$ on both sides, we have

$$\langle \mathbf{H}, \mathbf{B}_N\rangle \succeq \Big\langle \mathbf{H}, \prod_{t=1}^N (\mathcal{I}-\gamma_t\widetilde{\mathcal{T}}_t)\circ\mathbf{B}_0\Big\rangle + e^{-4}\beta\langle \mathbf{H}_{k^\dagger:\infty}, \mathbf{B}_0\rangle \cdot \underbrace{\Big\langle \mathbf{H}, \sum_{t=1}^N \gamma_t^2 \prod_{i=t+1}^N (\mathcal{I}-\gamma_i\widetilde{\mathcal{T}}_i)\circ\mathbf{H}\Big\rangle}_{(*)}.$$

Regarding $(*)$, we can define the function $f(x)$ as did (38), it is clear that $(*) = \sum_i f(\lambda_i)$ so that the results of Lemma E.2 can be directly applied. Moreover, note that

$$\prod_{t=1}^N (1-\gamma_t x)^2 \ge (1-\gamma x)^{2s}\cdot e^{-4\gamma x\sum_{t=1}^{N-s} t^{-a}} \ge \begin{cases} (1-\gamma x)^{2s}\cdot e^{-4\gamma x N^{1-a}/(1-a)}, & 0\le a<1 \\ (1-\gamma x)^{2s}\cdot e^{-4\gamma x\log(N)}, & a = 1. \end{cases}$$

Then combining the above results and applying the fact that $\mathbf{B}_0 = (\mathbf{w}_0-\mathbf{w}^*)(\mathbf{w}_0-\mathbf{w}^*)^\top$, we have

- **Case 1:** $0 \le a < 1$. Let $k^* = \max\{k : \gamma_0\lambda_k \ge (1-a)/[2(N-s)^{a-1}]\}$ and $k^\dagger = \max\{k : \gamma_0\lambda_k \ge 1/(2s)\}$, we have

$$\langle \mathbf{H}, \mathbf{B}_N \rangle \ge \big\|(\mathbf{I}-\gamma\mathbf{H})^{s+2N^{1-a}/(1-a)} \cdot (\mathbf{w}_0 - \mathbf{w}^*)\big\|_{\mathbf{H}}^2 + e^{-4}\beta\langle \mathbf{H}_{k^\dagger:\infty}, \mathbf{B}_0\rangle \cdot \bigg(\sum_{i\le k^*} \frac{(1-a)\cdot\gamma_0\lambda_i}{N^a} + \sum_{k^*+1\le i\le k^\dagger} \frac{\gamma_0\lambda_i}{4e^2} + \sum_{i\ge k^\dagger+1} \frac{s\gamma_0^2\lambda_i^2}{2e^2}\bigg).$$

- **Case 2:** $a = 1$**.** Let $k^* = \max\{k : \gamma_0\lambda_k \geq 1/(2 + 2\log(N - s - 1))\}$ and $k^\dagger = \max\{k : \gamma_0\lambda_k \geq 1/(2s)\}$, we have

$$\langle \mathbf{H}, \mathbf{B}_N \rangle \geq \left\| (\mathbf{I} - \gamma\mathbf{H})^{s+2\log(N)} \cdot (\mathbf{w}_0 - \mathbf{w}^*) \right\|_{\mathbf{H}}^2 + e^{-4}\beta \langle \mathbf{H}_{k^\dagger:\infty}, \mathbf{B}_0 \rangle \cdot \left(\sum_{i \leq k^*} \frac{\gamma_0^2\lambda_i^2}{N^{4\gamma_0\lambda_i}} + \sum_{k^*+1 \leq i \leq k^\dagger} \frac{\gamma_0\lambda_i}{4e^2} + \sum_{i \geq k^\dagger+1} \frac{s\gamma_0^2\lambda_i^2}{2e^2} \right).$$

This completes the proof.

□

E.3. Proof of Theorem 4.3

Here we state the full version of Theorem 4.3 and provide its proof.

Theorem E.3 (A lower bound for poly-decaying stepsizes)**.** *Consider last iterate SGD with stepsize scheme* (4)*. Suppose Assumptions 3.1, 3.2B and 3.3' hold. Suppose* $\gamma_0 < 1/(4\lambda_1)$ *and* $s\gamma_0 \geq \sum_{t=s+1}^{N} \gamma_t$*, then for any constant* $a \in [0, 1]$*,*

$$\mathbb{E}[L(\mathbf{w}_N) - L(\mathbf{w}^*)] = \frac{1}{2}\texttt{BiasError} + \frac{1}{2}\texttt{VarianceError}.$$

Moreover:

- *If* $0 \leq a < 1$*, then*

$$\texttt{BiasError} \geq \left\| (\mathbf{I} - \gamma_0\mathbf{H})^{s + \frac{2N^{1-a}}{1-a}} \cdot (\mathbf{w}_0 - \mathbf{w}^*) \right\|_{\mathbf{H}}^2 + \frac{(1-a)^2\beta}{e^4} \cdot \|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{H}_{k^\dagger:\infty}}^2 \cdot \frac{d_{\texttt{eff}}}{N},$$

  and

$$\texttt{VarianceError} \geq (1-a)^2\sigma^2 \cdot \frac{d_{\texttt{eff}}}{N}.$$

  Here $k^* := \max\{k : \gamma_0\lambda_k \geq (1-a)/(2(N-s)^{1-a})\}$*,* $k^\dagger := \max\{k : \gamma_0\lambda_k \geq 1/(2s)\}$*, and the* effective dimension *is defined by*

$$d_{\texttt{eff}} := \sum_{i \leq k^*} \max\{N^{1-a}\gamma_0\lambda_i, \frac{a\log(N)}{16e}\} + \sum_{k^* < i \leq k^\dagger} \frac{N\gamma_0\lambda_i}{4e^2} + \sum_{i > k^\dagger} \frac{sN\gamma_0^2\lambda_i^2}{2e^2}.$$

- *If* $a = 1$*, then*

$$\texttt{BiasError} \geq \left\| (\mathbf{I} - \gamma_0\mathbf{H})^{s+2\log(N)} \cdot (\mathbf{w}_0 - \mathbf{w}^*) \right\|_{\mathbf{H}}^2 + \frac{\beta}{e^4} \cdot \|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{H}_{k^\dagger:\infty}}^2 \cdot \frac{d_{\texttt{eff}}}{N},$$

  and

$$\texttt{VarianceError} \geq \sigma^2 \cdot \frac{d_{\texttt{eff}}}{N}.$$

  Here $k^* := \max\{k : \gamma_0\lambda_k \geq 1/(2 + 2\log(N - s - 1))\}$ $k^\dagger := \max\{k : \gamma_0\lambda_k \geq 1/(2s)\}$*, and the* effective dimension *is defined by*

$$d_{\texttt{eff}} := \sum_{i \leq k^*} N^{1-4\gamma_0\lambda_i}\gamma_0^2\lambda_i^2 + \sum_{k^* < i \leq k^\dagger} \frac{N\gamma_0\lambda_i}{4e^2} + \sum_{i > k^\dagger} \frac{sN\gamma_0^2\lambda_i^2}{2e^2}.$$

Proof. The proof is a simple combination of Lemmas B.3, E.2, and E.1. □

E.4. Proof of Theorem 4.4

Proof. The proof will be focusing showing that the upper bound for geometrically decaying stepsize (Theorem 4.1) is smaller than the lower bound for polynomially decaying stepsize (Theorem 4.3) up to some constant factors.

First, note that $\mathbf{I}_{0:k^*}/\gamma_0 K \preceq \mathbf{H}_{0:k^*}$, if setting $k^* = \max\{k : \gamma_0\lambda_k \geq 1/K\}$, then the first term of the bias error upper bound in Theorem 4.3 can be further relaxed as

$$\begin{aligned}
&\frac{\big\|(\mathbf{I}-\gamma_0\mathbf{H})^{s+K}(\mathbf{w}_0-\mathbf{w}^*)\big\|^2_{\mathbf{I}_{0:k^*}}}{\gamma_0 K} + \big\|(\mathbf{I}-\gamma_0\mathbf{H})^{s+K}(\mathbf{w}_0-\mathbf{w}^*)\big\|^2_{\mathbf{H}_{k^*:\infty}} \\
&\leq \big\|(\mathbf{I}-\gamma_0\mathbf{H})^{s+K}(\mathbf{w}_0-\mathbf{w}^*)\big\|^2_{\mathbf{H}_{0:k^*}} + \big\|(\mathbf{I}-\gamma_0\mathbf{H})^{s+K}(\mathbf{w}_0-\mathbf{w}^*)\big\|^2_{\mathbf{H}_{k^*:\infty}} \\
&= \|(\mathbf{I}-\gamma\mathbf{H})^{s+K}(\mathbf{w}-\mathbf{w}^*)\|^2_{\mathbf{H}}.
\end{aligned}$$

Moreover, note that we have set $s = N/2$, it is clear that $s\gamma \geq \sum_{t=s+1}\gamma_t$ for polynomially decaying stepsize so that Theorem 4.3 holds. Then it can be shown that the length of each phase in SGD with geometrically decaying stepsize is $K = (N-s)/\log(N-s) = \Theta(N/\log(N)) = \omega(N^{1-a} \vee \log(N))$. Therefore, we have

$$\|(\mathbf{I}-\gamma\mathbf{H})^{s+K}(\mathbf{w}-\mathbf{w}^*)\|^2_{\mathbf{H}} \leq \begin{cases} \|(\mathbf{I}-\gamma\mathbf{H})^{s+\frac{2N^{1-a}}{1-a}}(\mathbf{w}-\mathbf{w}^*)\|^2_{\mathbf{H}}, & \text{for any constant } a \in [0,1); \\ \|(\mathbf{I}-\gamma\mathbf{H})^{s+2\log(N)}(\mathbf{w}-\mathbf{w}^*)\|^2_{\mathbf{H}}, & a = 1. \end{cases}$$

Note that the second term in the bias error bound has a quite similar form as the variance bound. Then we will consider the variance error and the results can be directly applied to the second term in the bias error bound. By looking at the upper bound in Theorems 4.1 and 4.3, we can compare the variance error along different dimensions separately.

Case 1: $a \in [0,1)$ For the case $a \in [0,1)$, we define $k_1^* = \max\{k : \gamma_0\lambda_k \geq (1-a)/[2(N-s)^{1-a}]\} = \max\{k : \gamma_0\lambda_k \geq \Theta(1/N^{1-a})\}$, $k_2^* = \max\{k : \gamma_0\lambda_k \geq 1/K\} = \max\{k : \gamma_0\lambda_k \geq \Theta(\log(N)/N)\}$, and $k^\dagger = \max\{k : \gamma_0\lambda_k \geq \Theta(1/N)\}$, we have

$$\texttt{VarianceError}_{\text{exp}} \lesssim \sigma^2 \cdot \bigg(\sum_{i \leq k_1^*} \frac{\log(N)}{N} + \sum_{k_1^* < i \leq k_2^*} \frac{\log(N)}{N} + \sum_{k_2^* < i \leq k^\dagger} \gamma_0\lambda_i + \sum_{i > k^\dagger} N\gamma_0^2\lambda_i^2 \bigg)$$

and

$$\texttt{VarianceError}_{\text{poly}} \gtrsim \sigma^2 \cdot \bigg(\sum_{i \leq k_1^*} \frac{\gamma_0\lambda_i}{N^a} \vee \frac{\log(N)}{N} + \sum_{k_1^* < i \leq k_2^*} \gamma_0\lambda_i + \sum_{k_2^* < i \leq k^\dagger} \gamma_0\lambda_i + \sum_{i > k^\dagger} N\gamma_0^2\lambda_i^2 \bigg).$$

Then it suffices to consider the case $k_1^* < i \leq k_2^*$. In particular, according to the definition of k_1^* and k_2^*, it is clear that for any $k_1^* < i \leq k_2^*$,

$$\gamma_0\lambda_i \geq \frac{1}{K} = \Theta\big(\log(N)/N\big).$$

This implies that $\texttt{VarianceError}_{\text{exp}} \lesssim \texttt{VarianceError}_{\text{poly}}$. Then we can go back to the second term of the bias error bounds, which have a similar formula of the variance error bound. Applying the definition $R(N) = (\|\mathbf{w}-\mathbf{w}^*\|^2_{\mathbf{I}_{0:k^\dagger}}/(\gamma_0 N) + \|\mathbf{w}-\mathbf{w}^*\|_{\mathbf{H}_{k^\dagger:\infty}})/\sigma^2$, we can conclude that

$$\mathbb{E}[L(\mathbf{w}_N^{\text{geo}}) - L(\mathbf{w}^*)] \leq C \cdot [1 + \log(N) \cdot R(N)] \cdot \mathbb{E}[L(\mathbf{w}_N^{\text{poly}}) - L(\mathbf{w}^*)]$$

Case 2: $a = 1$**.** Similarly, we can now define $k_1^* := \max\{k : \gamma_0\lambda_k \geq 1/(2 + 2\log(N-s-1))\} = \max\{k : \gamma_0\lambda_k \geq \Theta(1/\log(N))\}$ and get

$$\texttt{VarianceError}_{\text{poly}} \gtrsim \sigma^2 \cdot \bigg(\sum_{i \leq k_1^*} \frac{\gamma_0^2\lambda_i^2}{N^{4\gamma_0\lambda_i}} + \sum_{k_1^* < i \leq k_2^*} \gamma_0\lambda_i + \sum_{k_2^* < i \leq k^\dagger} \gamma_0\lambda_i + \sum_{i \geq k^\dagger + 1} N\gamma_0^2\lambda_i^2 \bigg).$$

Note that we have assumed $4\gamma_0\lambda_i < 1$, then we have for all $i \leq k_1^*$

$$\frac{\gamma_0^2\lambda_i^2}{N^{4\gamma_0\lambda_i}} \geq N^{-4\gamma_0\lambda_i}/\log^2(N) = \Omega(\log(N)/N).$$

Additionally, for any $k_1^* < i \leq k_2^*$, we also have

$$\gamma_0\lambda_i \geq \frac{1}{K} = \Theta\big(\log(N)/N\big).$$

Then combining the bias error and variance bounds, we can also conclude that

$$\mathbb{E}[L(\mathbf{w}_N^{\mathrm{geo}}) - L(\mathbf{w}^*)] \leq C \cdot [1 + \log(N) \cdot R(N)] \cdot \mathbb{E}[L(\mathbf{w}_N^{\mathrm{poly}}) - L(\mathbf{w}^*)]$$

where $R(N) = (\|\mathbf{w} - \mathbf{w}^*\|^2_{\mathbf{I}_{0:k^\dagger}}/(\gamma_0 N) + \|\mathbf{w} - \mathbf{w}^*\|_{\mathbf{H}_{k^\dagger:\infty}})/\sigma^2$. This completes the proof. □