Journal of Machine Learning Research 24 (2023) 1-58 Submitted 11/21; Revised 9/22; Published 10/23

Benign Overfitting of Constant-Stepsize SGD for Linear Regression

Difan Zou* DZOU@CS.HKU.HK
Department of Computer Science & Institute of Data Science
The University of Hong Kong

Jingfeng Wu* UUUJF@BERKELEY.EDU
Simons Institute
University of California, Berkeley

Vladimir Braverman VB21@RICE.EDU
Department of Computer Science
Rice University

Quanquan Gu QGU@CS.UCLA.EDU
Department of Computer Science
University of California, Los Angeles

Sham M. Kakade SHAM@SEAS.HARVARD.EDU
Department of Computer Science & Department of Statistics
Harvard University

Editor: Karthik Sridharan

Abstract

There is an increasing realization that algorithmic inductive biases are central in preventing overfitting; empirically, we often see a *benign overfitting* phenomenon in overparameterized settings for natural learning algorithms, such as stochastic gradient descent (SGD), where little to no *explicit* regularization has been employed. This work considers this issue in arguably the most basic setting: *constant-stepsize SGD* (with iterate averaging or tail averaging) for linear regression in the overparameterized regime. Our main result provides a sharp excess risk bound, stated in terms of the full eigenspectrum of the data covariance matrix, that reveals a bias-variance decomposition characterizing when generalization is possible: (i) the variance bound is characterized in terms of an *effective dimension* (specific for SGD) and (ii) the bias bound provides a sharp geometric characterization in terms of the location of the initial iterate (and how it aligns with the data covariance matrix). More specifically, for SGD with iterate averaging, we demonstrate the sharpness of the established excess risk bound by proving a matching lower bound (up to constant factors). For SGD with tail averaging, we show its advantage over SGD with iterate averaging by proving a better excess risk bound together with a nearly matching lower bound. Moreover, we reflect on a number of notable differences between the algorithmic regularization afforded by (unregularized) SGD in comparison to ordinary least squares (minimum-norm interpolation) and ridge regression. Experimental results on synthetic data corroborate our theoretical findings [1].

[*] Equal Contribution. The work was done when DZ was a PhD student at UCLA, and JW was a PhD student at Johns Hopkins University.

[1] A short version is accepted at the *34th Annual Conference on Learning Theory* (COLT 2021).

1. Introduction

A widely observed and yet still striking phenomenon is that modern machine learning models (e.g., deep neural networks) trained by stochastic gradient descent often generalize while also achieving near-zero training error (i.e., despite being overparameterized and under-regularized [2]. See Belkin et al. (2020) for further discussion.). There is reason to believe that characterizing these effects even in conceptually simpler (e.g. linear model) settings will also help our understanding of more complex settings, because many high dimensional effects are also observed even in simple linear models. For example, this *benign overfitting* effect is also observed for the ordinary least square (OLS) estimator, where it is observed that OLS generalizes in the overparameterized regime (Bartlett et al., 2020).

For OLS in particular, the recent work of Bartlett et al. (2020) established non-asymptotic generalization guarantees of the *minimum-norm interpolator* for overparameterized linear regression (the minimum-norm solution that *perfectly fits* the training samples (Zhang et al., 2016; Bartlett et al., 2020)). More generally, there is a growing body of work studying generalization in basic linear models in the overparameterized regime (Nakkiran et al., 2019; Bartlett et al., 2020; Belkin et al., 2020; Hastie et al., 2019; Tsigler and Bartlett, 2020; Muthukumar et al., 2020; Chatterji and Long, 2020; Nakkiran et al., 2020). In contrast, for *stochastic gradient descent* (SGD) for least squares regression, the algorithmic aspects of generalization are far less well understood, where we lack a sharp characterization of it and when benign overfitting occurs (in other words, achieving diminishing generalization error). This is the focus of this work.

With regards to SGD in the classical underparameterized regime, the seminal work of Polyak and Juditsky (1992) showed that iterate averaged SGD achieves, in the limit as the sample size goes to infinity, the statistically optimal rate, even up to problem dependent constant factors [3]; this optimality crucially relies on the dimension being held finite, along with regularity assumptions that make the problem locally strongly quadratic. For the case of *finite* dimensional, linear regression, there are a number of more modern proofs which provide finite, non-asymptotic rates (Défossez and Bach, 2015; Bach and Moulines, 2013; Dieuleveut et al., 2017; Jain et al., 2017b,a). Besides, for smooth and strongly convex problems, Rakhlin et al. (2012) showed that the iterate averaged SGD can achieve the minimax optimal rate. With regards to the overparameterized regime, there is far less work (Dieuleveut and Bach, 2015; Berthier et al., 2020) being notable exceptions. (See Section 3 for further discussion on these related works.)

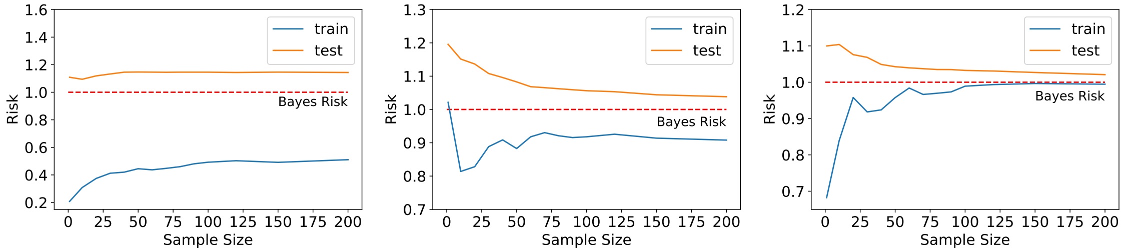
SGD for linear regression. The classical linear regression problem of interest is:

$$\min_{\mathbf{w}} L(\mathbf{w}), \text{ where } L(\mathbf{w}) = \frac{1}{2}\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\big[(y - \langle \mathbf{w}, \mathbf{x}\rangle)^2\big], \tag{1.1}$$

where $\mathbf{x} \in \mathcal{H}$, is the feature vector, where, $\mathcal{H}$ is some (finite d-dimensional or countably infinite dimensional) Hilbert space; $y \in \mathbb{R}$ is the response; $\mathcal{D}$ is an unknown distribution

[2] By "under-regularized", we mean that the empirical training loss is near to 0, such as with OLS when $N \gg d$.

[3] Polyak and Juditsky (1992) provided a stronger distributional limit theorem showing that the distribution of the averaged iterate (provided by SGD) precisely matches the distribution of the empirical risk minimizer.



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

Figure 1: Benign overfitting of SGD for linear regression. The plots show the training and test risks achieved by SGD (constant stepsize, iterate averaging) for least square problem instances (the spectrum of $\mathbf{H}$, i.e., $\{\lambda_i\}$ is specified under each subfigure). The problem dimension is $d = 2000$ and the variance of model noise is $\sigma^2 = 1$ (hence the Bayes risk is 1). The plots are averaged over 20 independent runs. In (a), SGD overfits the training sample (achieving a training risk smaller than the Bayes risk) but generalizes poorly. In (b), SGD overfits the training sample and generalizes well, which exhibits the benign overfitting phenomenon. In (c), SGD generalizes on test samples and tends to forget training samples, which indicates a regularization effect of SGD. See Section 6 for more details.

over $\mathbf{x}$ and y; and $\mathbf{w} \in \mathcal{H}$ is the weight vector to be optimized. We consider the stochastic approximation approach using constant stepsize SGD, with iterate averaging: at each iteration t, an i.i.d. example $(\mathbf{x}_t, y_t) \sim \mathcal{D}$ is observed, and the weight is updated according to SGD as follows:

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \gamma \left(y_t - \langle \mathbf{w}_{t-1}, \mathbf{x}_t\rangle\right) \mathbf{x}_t, \qquad t = 1, \dots, N, \tag{1.2}$$

where $\gamma > 0$ is a constant stepsize, N is the number of samples observed, and the weights are initialized at $\mathbf{w}_0 \in \mathcal{H}$. The final output will be the average of the iterates:

$$\overline{\mathbf{w}}_N := \frac{1}{N} \sum_{t=0}^{N-1} \mathbf{w}_t.$$

In the underparameterized setting with finite dimension d ($d \ll N$), as mentioned earlier (also see Section 3), a rich body of work has established that $\overline{\mathbf{w}}_N$ enjoys the optimal risk (up to constant factors) of $\mathcal{O}\left(d\sigma^2/N\right)$, for sufficiently large N. The focus of this work is on the overparameterized regime, where $d \gg N$ (or possibly countably infinite).

Benign overfitting occurs in SGD for linear regression. Perhaps quite surprisingly, the benign overfitting phenomenon, i.e., a predictor that fits training data very well but still generalizes, happens for SGD (with constant stepsize and iterate averaging) even for the simple, overparameterized linear regression. This is empirically verified in Figure 1, where we see in Figure 1 (b) that, SGD overfits the training sample (achieving a training risk much lower than the Bayes risk) but still generalizes on the test sample (the test risk is vanishing). Understanding this phenomenon theoretically is one of the central goals of this work.

Our contributions. Our main result can be viewed as a counterpart to the classical analysis of iterate averaged SGD to the overparameterized regime for linear regression: we provide a sharp excess risk bound showing how (unregularized) SGD can generalize even in the infinite-dimensional setting. Our bound is stated in a general manner, in terms of the full eigenspectrum of the data covariance matrix along with a functional dependency on the initial iterate; our lower bound shows our characterization is tight. As a corollary, we see how the benign-overfitting phenomenon can be observed for SGD, provided certain spectrum decay conditions on the data covariance are met. We also extend our results to SGD with tail-averaging (Polyak and Juditsky, 1992; Jain et al., 2017a,b), where we run SGD for s iterations and then take average over the subsequent N iterates as the output. (see Section 5 for more details.)

Some additional notable contributions are:

1. The sharpness of our bounds permits us to make comparisons to OLS (the minimum-norm interpolator) and ridge regression. Notably, in a contrast to the variance of OLS (Bartlett et al., 2020), the variance contribution to SGD is well behaved under substantially weaker assumptions on the spectrum of the data covariance. This shows how inductive bias of SGD, in comparison to the minimum-norm interpolator, can lead to better generalization with no regularization. We also constrast our results to ridge regression based on the recent work by Tsigler and Bartlett (2020).

2. One notable aspect of our work is a sharp characterization of a "bias process" in SGD. In particular, consider the special case where $y = \mathbf{w}^\star \cdot \mathbf{x}$ (with probability one), for some $\mathbf{w}^\star$. Here, SGD still differs from gradient descent on $L(\mathbf{w})$. Our characterization gives a novel characterization of how the variance in this process contributes to the final excess risk bound.

3. From a technical standpoint, our work develops new proof techniques for iterate averaged SGD. Our analysis tools are based on the operator view of averaged SGD (Dieuleveut and Bach, 2015; Jain et al., 2017b,a). A core idea in the proof is in connecting the finite sample (infinite dimensional) covariance matrices of the variance and bias stochastic processes to those of their corresponding (asymptotic) stationary covariance matrices — an idea that was introduced in Jain et al. (2017a) for the finite dimensional, variance analysis.

Notation. We use lower case letters to denote scalars, and we use lower and upper case bold face letters to denote vectors and matrices respectively. For a vector $\mathbf{x} \in \mathcal{H}$, $\|\mathbf{x}\|_2$ denotes the norm in the Hilbert space $\mathcal{H}$, and $\mathbf{x}[i]$ denotes the i-th coordinate of $\mathbf{x}$. For a matrix $\mathbf{M}$, its spectral norm is denoted by $\|\mathbf{M}\|_2$. For a PSD matrix $\mathbf{A}$, define $\|\mathbf{x}\|_{\mathbf{A}}^2 := \mathbf{x}^\top \mathbf{A}\mathbf{x}$.

2. Main Results

We now provide matching (upto absolute constants) upper and lower excess risk bounds for iterate averaged SGD. We then compare these rates to those of OLS and ridge regression, where we see striking similarities and notable differences.

2.1 Benign Overfitting of SGD

We first introduce relevant notation and our assumptions. Our first assumption is mild regularity conditions on the moments of the data distribution.

Assumption 2.1 (Regularity conditions) *Assume* $\mathbb{E}[\mathbf{x}\mathbf{x}^\top]$, $\mathbb{E}[\mathbf{x}\otimes\mathbf{x}\otimes\mathbf{x}\otimes\mathbf{x}]$, *and* $\mathbb{E}[y^2]$ *exist and are all finite. Furthermore, denote the second moment of* $\mathbf{x}$ *by*

$$\mathbf{H} := \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}\mathbf{x}^\top],$$

and suppose that $\mathrm{tr}(\mathbf{H})$ *is finite. For convenience, we assume that* $\mathbf{H}$ *is strictly positive definite and that* $L(\mathbf{w})$ *admits a unique global optimum, which we denote by* $\mathbf{w}^* := \mathrm{argmin}_{\mathbf{w}} L(\mathbf{w})$*.* [4]

Our second assumption is on the behavior of the fourth moment, when viewed as a linear operator on PSD matrices:

Assumption 2.2 (Fourth moment condition) *Assume there exists a positive constant* $\alpha > 0$*, such that for any PSD matrix* $\mathbf{A}$[5]*, it holds that*

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}\mathbf{x}^\top\mathbf{A}\mathbf{x}\mathbf{x}^\top] \preceq \alpha\,\mathrm{tr}(\mathbf{H}\mathbf{A})\mathbf{H}.$$

For Gaussian distributions, it suffices to take $\alpha = 3$*. Furthermore, it is worth noting that this assumption is implied if the distribution over* $\mathbf{H}^{-\frac{1}{2}}\mathbf{x}$ *has sub-Gaussian tails (see Lemma A.1 in the Appendix for a precise claim). Also, it is not difficult to verify that* $\alpha \geq 1$*.*[6]

Assuming sub-Gaussian tails over $\mathbf{H}^{-\frac{1}{2}}\mathbf{x}$ is a standard assumption in linear regression analysis (e.g., Hsu et al. (2014); Bartlett et al. (2020); Tsigler and Bartlett (2020)), which is also necessary to give a sharp excess risk bound for the minimum-norm interpolator (Bartlett et al., 2020). However, as mentioned above, this assumption is strictly stronger than Assumption 2.2. The assumption is somewhat stronger than what is often assumed for iterate averaged SGD in the underparameterized regime (e.g., Bach and Moulines 2013; Jain et al. 2017b) (see Section 3 for further discussion). Additionally, we also remark that Assumption 2.2 can be further relaxed to that we only require $\mathbf{A}$ is PSD and commutable with $\mathbf{H}$, rather than all PSD matrix $\mathbf{A}$ (see Section 7 for more details).

Our next assumption is a noise condition, where it is helpful to interpret $y - \langle\mathbf{w}^*, \mathbf{x}\rangle$ as the additive noise. Observe that the first order optimality conditions on $\mathbf{w}^*$ imply $\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - \langle\mathbf{w}^*, \mathbf{x}\rangle)\mathbf{x}] = \nabla L(\mathbf{w}^*) = \mathbf{0}$.

[4] This is not necessary. In the case where $\mathbf{H}$ has eigenvalues which are 0, we could instead choose $\mathbf{w}^*$ to be the minimum norm vector in the set $\mathrm{argmin}_{\mathbf{w}} L(\mathbf{w})$, and our results would hold for this choice of $\mathbf{w}^\star$. For example, see Schölkopf et al. (2002) for a rigorous treatment of working in a reproducing kernel Hilbert space.

[5] This assumption can be relaxed into: for any PSD matrix $\mathbf{A}$ *that commutes with* $\mathbf{H}$, it holds that $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}\mathbf{x}^\top\mathbf{A}\mathbf{x}\mathbf{x}^\top] \preceq \alpha\,\mathrm{tr}(\mathbf{H}\mathbf{A})\mathbf{H}$. The presented analyzing technique is ready to be modified to cooperate with the relaxed assumption with the observation that the fourth moment operator is linear and self-adjoint. Similar relaxation applies to Assumption 2.4 as well.

[6] This is due to that the square of the second moment is less than the fourth moment.

Assumption 2.3 (Noise condition) *Suppose that:*

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

exist and are finite. Note that $\mathbf{\Sigma}$ is the covariance matrix of the gradient noise at $\mathbf{w}^\star$.

This assumption places a rather weak requirement on the additive noise (due to that it permits model mis-specification) and is often made in the average SGD literature (e.g., Bach and Moulines 2013; Dieuleveut et al. 2017). Observe that for *well-specified models*, where

$$y = \langle \mathbf{w}^\star, \mathbf{x}\rangle + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2_{\text{noise}}), \tag{2.1}$$

we have that $\mathbf{\Sigma} = \sigma^2_{\text{noise}}\mathbf{H}$ and so $\sigma^2 = \sigma^2_{\text{noise}}$.

We would like to further point out that Assumptions 2.1, 2.2 and 2.3 are applicable to both overparameterized and underparameterized regimes since they are made on the population covariance matrix and data distribution, which are independent of the training sample size.

Before we present our main theorem, a few further definitions are in order: denote the eigendecomposition of the Hessian as $\mathbf{H} = \sum_i \lambda_i \mathbf{v}_i\mathbf{v}_i^\top$, where $\{\lambda_i\}_{i=1}^\infty$ are the eigenvalues of $\mathbf{H}$ sorted in non-increasing order and $\mathbf{v}_i$'s are the corresponding eigenvectors. We then denote:

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

Similarly we denote $\mathbf{I}_{0:k} := \sum_{i=1}^k \mathbf{v}_i\mathbf{v}_i^\top$ and $\mathbf{I}_{k:\infty} := \sum_{i>k}\mathbf{v}_i\mathbf{v}_i^\top$. By the above definitions, we know

$$\|\mathbf{w}\|^2_{\mathbf{H}^{-1}_{0:k}} = \sum_{i\le k}\frac{(\mathbf{v}_i^\top\mathbf{w})^2}{\lambda_i}, \quad \|\mathbf{w}\|^2_{\mathbf{H}_{k:\infty}} = \sum_{i>k}\lambda_i(\mathbf{v}_i^\top\mathbf{w})^2,$$

where we have slightly abused notation in that $\mathbf{H}^{-1}_{0:k}$ denotes a pseudo-inverse.

We now present our main theorem:

Theorem 2.1 (Benign overfitting of SGD) *Suppose Assumptions 2.1-2.3 hold and that the stepsize is set so that $\gamma < 1/(\alpha\sum_i\lambda_i)$. Then the excess risk can be upper bounded as follows,*

$$\mathbb{E}[L(\overline{\mathbf{w}}_N)] - L(\mathbf{w}^*) \le 2\cdot\text{EffectiveBias} + 2\cdot\text{EffectiveVar},$$

where

$$\begin{aligned}
\text{EffectiveBias} &= \frac{1}{\gamma^2N^2}\cdot\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k^*}} + \|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}},\\
\text{EffectiveVar} &= \frac{2\alpha\big(\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{I}_{0:k^*}} + N\gamma\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}}\big)}{N\gamma(1-\gamma\alpha\sum_i\lambda_i)}\cdot\left(\frac{k^*}{N} + N\gamma^2\sum_{i>k^*}\lambda_i^2\right)\\
&\quad + \frac{\sigma^2}{1-\gamma\alpha\sum_i\lambda_i}\cdot\left(\frac{k^*}{N} + N\gamma^2\sum_{i>k^*}\lambda_i^2\right)
\end{aligned}$$

with $k^* = \max\{k : \lambda_k \ge \frac{1}{\gamma N}\}$.

The interpretation is as follows: the "effective bias" precisely corresponds to the rate of convergence had we run gradient descent directly on $L(\mathbf{w})$ (i.e., where the latter has no variance due to sampling). The "effective variance" error stems from both the additive noise $y - \langle \mathbf{w}^*, \mathbf{x} \rangle$, i.e., the second term of the EffectiveVariance error, along with that even if there was no additive noise (i.e. $y - \langle \mathbf{w}^*, \mathbf{x} \rangle = 0$ with probability one), i.e., the first term of the EffectiveVariance error, then SGD would still not be equivalent to GD. The cut-off index k^*, which we refer to as the "effective dimension", plays a pivotal role in the excess risk bound, which separates the entire space into a k^*-dimensional "head" subspace where the bias error decays more quickly than that of the bias error in the complement "tail" subspace. To obtain a vanishing bound, the effective dimension k^* must be $\mathrm{o}\left(N\right)$ and the tail summation $\sum_{i>k^*} \lambda_i^2$ must be $\mathrm{o}\left(1/N\right)$.

Additionally, we would like to point out that our bounds are valid (i.e., have finite upper bound) as long as the global optimum $\mathbf{w}^*$ and the initialization $\mathbf{w}_0$ satisfy $\|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{H}}^2 < \infty$, which is equivalent to have a finite initial population risk, i.e., $L_{\mathcal{D}}(\mathbf{w}_0) \le \infty$. As a comparison, prior works also assumed the *source conditions* on $\mathbf{w}^*$: Dieuleveut and Bach (2015) assumed $\|\mathbf{H}^{-\alpha}(\mathbf{w}_0 - \mathbf{w}^*)\|_2 < \infty$ for $\alpha \ge -1/2$, which is the same as our implicit condition on $\mathbf{w}^*$; Berthier et al. (2020) assumed $\|\mathbf{H}^{-\alpha}(\mathbf{w}_0 - \mathbf{w}^*)\|_2 < \infty$ for $\alpha \ge 0$, which is stronger than that in Dieuleveut and Bach (2015) and our paper.

In terms of constant factors, the above bound can be improved by a factor of 2 in the effective bias-variance decomposition (see (4.6)). We now turn to lower bounds.

A lower bound. We first introduce the following assumption that states a lower bound on the fourth moment.

Assumption 2.4 (Fourth moment condition, lower bound) *Assume there exists a constant $\beta \ge 0$, such that for any PSD matrix $\mathbf{A}$, it holds that*

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}\mathbf{x}^\top \mathbf{A}\mathbf{x}\mathbf{x}^\top] - \mathbf{H}\mathbf{A}\mathbf{H} \succeq \beta\, \mathrm{tr}(\mathbf{H}\mathbf{A})\mathbf{H}.$$

For Gaussian distributions, it suffices to take $\beta = 1$.

The following lower bound shows that when the noise is well-specified our upper bound is not improvable except for absolute constants.

Theorem 2.2 (Excess risk lower bound) *Suppose $N \ge 500$. For any well-specified data distribution $\mathcal{D}$ (see (2.1)) that also satisfies Assumptions 2.1 and 2.4, for any stepsize such that $\gamma < 1/\lambda_1$, we have that:*

$$\begin{aligned}\mathbb{E}[L(\overline{\mathbf{w}}_N)] - L(\mathbf{w}^*) \ge \frac{1}{100\gamma^2 N^2} \cdot \|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k^*}} + \frac{1}{100} \cdot \|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}} \\ + \frac{\beta\left(\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{I}_{0:k^*}} + N\gamma\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}}\right)}{1000N\gamma} \cdot \left(\frac{k^*}{N} + N\gamma^2 \sum_{i>k^*} \lambda_i^2\right) \\ + \frac{\sigma^2_{\mathrm{noise}}}{50} \cdot \left(\frac{k^*}{N} + N\gamma^2 \sum_{i>k^*} \lambda_i^2\right)\end{aligned}$$

with $k^* = \max\{k : \lambda_k \ge \frac{1}{N\gamma}\}$.

It is worth noting that the lower bound in Theorem 2.2 requires the data distribution to be well-specified, which is stronger than the noise condition (Assumption 2.3) made in the upper bound results (Theorem 2.1). Similar to the upper bound stated in Theorem 2.1, the first two terms represent the EffectiveBias and the last two terms represent the EffectiveVariance, in which the third and last terms are contributed by the model noise and variance in SGD. Our upper bound matches our lower bound up to absolute constants, which indicates the obtained rates are tight, at least for Gaussian data distribution with well-specified noise.

Special cases. It is instructive to consider a few special cases of Theorem 2.1. We first show the result for SGD with large stepsizes.

Corollary 2.1 (Benign overfitting with large stepsizes) *Suppose Assumptions 2.1-2.3 hold and that the stepsize is set to* $\gamma = 1/(2\alpha\sum_i \lambda_i)$*. Then*

$$\mathrm{EffectiveBias} = \frac{4\alpha^2(\sum_i \lambda_i)^2}{N^2}\cdot \|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k^*}} + \|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}}$$

$$\mathrm{EffectiveVar} = \big(2\sigma^2 + 4\alpha^2\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}}\big)\cdot\left(\frac{k^*}{N} + \frac{N\sum_{i>k^*}\lambda_i^2}{4\alpha^2(\sum_i \lambda_i)^2}\right),$$

where $k^* = \max\{k : \lambda_k \ge \frac{2\alpha\sum_i\lambda_i}{N}\}$*.*

Note that the bias error decays at different rates in different subspaces. Crudely, in the "head" eigenspace (spanned by the eigenvectors corresponding to large eigenvalues) the bias error decays in a faster $\mathcal{O}\left(1/N^2\right)$ rate (though there is weighting of λ_i in the head), while in the remaining "tail" eigenspace, the bias error decays at a slower $\mathcal{O}\left(1/N\right)$ rate (due to that all the eigenvalues in the tail are less than $\mathcal{O}\left(1/N\right)$). The following corollary provides a crude bias bound, showing that bias never decays more slowly than $\mathcal{O}\left(1/N\right)$.

Corollary 2.2 (Crude bias-bound) *Suppose Assumptions 2.1-2.3 hold and that the stepsize is set to* $\gamma = 1/(2\alpha\sum_i \lambda_i)$*. Then*

$$\mathbb{E}[L(\overline{\mathbf{w}}_N)] - L(\mathbf{w}^*) \le \frac{8\alpha\|\mathbf{w}_0 - \mathbf{w}^*\|_2^2\cdot\sum_i\lambda_i}{N} + 4\sigma^2\cdot\left(\frac{k^*}{N} + \frac{N\sum_{i>k^*}\lambda_i^2}{4\alpha^2(\sum_i \lambda_i)^2}\right),$$

where $k^* = \max\{k : \lambda_k \ge \frac{2\alpha\sum_i\lambda_i}{N}\}$*.*

Theorems 2.1 and 2.2 suggests that the excess risk achieved by SGD depends on the spectrum of the covariance matrix. The following corollary gives examples of data spectrum such that the excess risk is diminishing.

Corollary 2.3 (Example data distributions) *Under the same conditions as Theorem 2.1, suppose* $\|\mathbf{w}_0 - \mathbf{w}^*\|_2$ *is bounded.*

1. *For* $\mathbf{H}\in\mathbb{R}^{d\times d}$*, let* $s = N^r$ *and* $d = N^q$ *for some positive constants* $0 < r \le 1$ *and* $q \ge 1$*. If the spectrum of* $\mathbf{H}$ *satisfies*

$$\lambda_k = \begin{cases} 1/s, & k \le s,\\ 1/(d-s), & s+1 \le k \le d,\end{cases}$$

then $\mathbb{E}[L(\overline{\mathbf{w}}_N)] - L(\mathbf{w}^*) = \mathcal{O}\left(N^{r-1} + N^{1-q}\right)$*.*

2. *If the spectrum of* $\mathbf{H}$ *satisfies* $\lambda_k = k^{-(1+r)}$ *for some* $r > 0$*, then* $\mathbb{E}[L(\overline{\mathbf{w}}_N)] - L(\mathbf{w}^*) = \mathcal{O}\left(N^{-r/(1+r)}\right)$*.*

3. *If the spectrum of* $\mathbf{H}$ *satisfies* $\lambda_k = k^{-1}\log^{-\beta}(k+1)$ *for some* $\beta > 1$*, then* $\mathbb{E}[L(\overline{\mathbf{w}}_N)] - L(\mathbf{w}^*) = \mathcal{O}\left(\log^{-\beta}(N)\right)$*.*

4. *If the spectrum of* $\mathbf{H}$ *satisfies* $\lambda_k = e^{-k}$*, then* $\mathbb{E}[L(\overline{\mathbf{w}}_N)] - L(\mathbf{w}^*) = \mathcal{O}\left(\log(N)/N\right)$*.*

2.2 Comparisons to OLS and Ridge Regression

We now compare these rates to those obtained by OLS or ridge regression.

SGD vs. minimum-norm solution of OLS. In a more restrictive setting where the whitened data $\mathbf{H}^{-1/2}\mathbf{x}$ has a sub-Gaussian tail and independent components, Bartlett et al. (2020) proved that the minimum ℓ_2 norm interpolator for the linear regression problem on N training examples, denoted by $\widehat{\mathbf{w}}_N$, gives the following excess risk lower bound:

$$\mathbb{E}[L(\widehat{\mathbf{w}}_N)] - L(\mathbf{w}^*) \ge c\sigma^2\bigg(\frac{k^\star}{N} + \frac{N\sum_{i>k^\star}\lambda_i^2}{(\sum_{i>k^*}\lambda_i)^2}\bigg),$$

where c is an absolute constant, σ^2 is the variance of model noise, and $k^\star = \min\{k \ge 0 : \sum_{i>k}\lambda_i/\lambda_{k+1} \ge bN\}$ for some constant $b > 0$. It is clear that in order to achieve benign overfitting, one needs to ensure that $k^\star = \mathrm{o}(N)$ and $\sum_{i>k^\star}\lambda_i^2/(\sum_{i>k^\star}\lambda_i)^2 = \mathrm{o}(1/N)$. The first requirement prefers slow decaying rate of the data spectrum since one hopes to get a large $\sum_{i>k}\lambda_i/\lambda_{k+1}$ for small k. On the contrary, the second requirement suggests that the spectrum should decay fast enough since we need to ensure that the tail summation $\sum_{i>k^\star}\lambda_i^2$ is small. Consequently, as shown in Theorem 6 in Bartlett et al. (2020), if the data spectrum decays in a rate $\lambda_k = k^{-\alpha}\log^{-\beta}(k+1)$, the minimum ℓ_2-norm interpolator can achieve vanishing excess risk only when $\alpha = 1$ and $\beta \ge 1$. In contrast, our results show that SGD can achieve vanishing excess risk for any $\alpha > 1$ and $\beta \ge 0$ (as well as the case of $\alpha = 1$ and $\beta > 1$, see Corollary 2.3 for details) since a fast decaying spectrum can ensure both small k^* (the effective dimension) and small tail summation $\sum_{i>k^\star}\lambda_i^2$.

SGD vs. ridge regression. Tsigler and Bartlett (2020) show that the ridge regression estimator, denoted by $\widehat{\mathbf{w}}_N^\lambda$, has the following lower bound on the excess risk:

$$\mathbb{E}[L(\widehat{\mathbf{w}}_N^\lambda)] - L(\mathbf{w}^*) \ge \max_k\Bigg\{c_1\sum_i\frac{\lambda_i\mathbf{w}^*[i]^2}{(1+\lambda_i/(\lambda_{k+1}\rho_k))^2} + \frac{c_2}{n}\sum_i\min\bigg(1, \frac{\lambda_i^2}{\lambda_{k+1}^2(\rho_k+2)^2}\bigg)\Bigg\},$$

where λ is the regularization parameter, c_1 and c_2 are absolute constants and $\rho_k = \big(\lambda + \sum_{i>k}\lambda_i\big)/(N\lambda_{k+1})$. Tsigler and Bartlett (2020) further show that the lower bound nearly matches the following upper bound of the excess risk:

$$\begin{aligned}\mathbb{E}[L(\widehat{\mathbf{w}}_N^\lambda)] - L(\mathbf{w}^*) \le c_1'\bigg(&\|\mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k^\star}}\cdot\bigg(\frac{\lambda+\sum_{i>k}\lambda_i}{N}\bigg)^2 + \|\mathbf{w}^*\|^2_{\mathbf{H}_{k^\star:\infty}}\bigg)\\ &+ c_2'\sigma^2\bigg(\frac{k^\star}{N} + \frac{N\sum_{i>k^\star}\lambda_i^2}{(\lambda+\sum_{i>k^*}\lambda_i)^2}\bigg),\end{aligned}$$

where c_1' and c_2' are absolute constants, and $k^\star = \min\{k \ge 0 : (\sum_{i>k}\lambda_i + \lambda)/\lambda_{k+1} \ge bN\}$ for some constant $b > 0$. Comparing this to Corollary 2.1 suggests that SGD (using a constant stepsize with iterate averaging) may exhibit an implicit regularization effect that performs comparably to ridge regression with a constant regularization parameter (here we assume that $\mathrm{tr}(\mathbf{H})$ is of a constant order). A more direct problem-dependent comparison (e.g., consider the optimal learning rate for SGD and optimal λ for ridge regression) is a fruitful direction of further study, to more accurately gauge the differences between the implicit regularization afforded by SGD and the explicit regularization of ridge regression.

3. Further Related Work

We first discuss the work on iterate averaging in the finite dimensional case before turning to the overparameterized regime. In the underparameterized regime, where d is assumed to be finite, the behavior of constant stepsize SGD with iterate average or tail average has been well investigated from the perspective of the *bias-variance decomposition* (Défossez and Bach, 2015; Dieuleveut et al., 2017; Lakshminarayanan and Szepesvari, 2018; Jain et al., 2017a,b). For iterate averaging from the beginning, Défossez and Bach (2015); Dieuleveut et al. (2017) show a $\mathcal{O}\left(1/N^2\right)$ convergence rate for the bias error and a $\mathcal{O}\left(d/N\right)$ convergence rate for the variance error, where N is the number of observed samples and d is the number of parameters. The bias error rate can be further improved by considering averaging only the tail iterates (Jain et al., 2017a,b, 2018), provided that the minimal eigenvalue of $\mathbf{H}$ is bounded away from 0. We note that the work in Jain et al. (2017a,b, 2018) also give the optimal rates with model misspecification. These results all have dimension factors d and do not apply to the overparameterized regime, though our results recover the finite dimensional case (and the results for delayed tail averaging from Jain et al. (2017a,b) can be applied here for the bias term). We further develop on the proof techniques in Jain et al. (2017a), where we use properties of asymptotic stationary distributions for the purposes of finite sample size analysis.

Another notable difference in our work is that Assumption 2.2 (which is implied by sub-Gaussianity, see Lemma A.1) is somewhat stronger than what is often assumed for iterate average SGD analysis, where $\mathbb{E}[\mathbf{x}\mathbf{x}^\top\mathbf{x}\mathbf{x}^\top] \preceq R^2\mathbf{H}$, as adopted in Bach and Moulines (2013); Défossez and Bach (2015); Dieuleveut et al. (2017); Jain et al. (2017a,b). Our assumption implies an R^2 bound with $R^2 = \alpha\,\mathrm{tr}(\mathbf{H})$. In terms of analysis, we note that our variance analysis only relies on an R^2 condition, while our bias analysis relies on our stronger sub-Gaussianity-like assumption.

Compared with Dieuleveut and Bach (2015), our bounds apply to least square instances with *any* data covaraince spectrum (under Assumption 2.2) and are established in an **instance-wise** manner, while the theoretical results in Dieuleveut and Bach (2015) are developed under the certain capacity conditions (e.g., $\lambda_i \le s^2/i^\alpha$ for some $\alpha > 1$ and $s > 0$, see A3 in Dieuleveut and Bach (2015)). Therefore, the bounds proved in Dieuleveut and Bach (2015) are not instance-wise, i.e., their bounds will be the same for two different problem instances that satisfy the capacity conditions with the same s and α. In comparison with Berthier et al. (2020), their bounds rely on a weaker fourth moment assumption, but rely on a stronger source condition on the ground truth parameter $\mathbf{w}^*$ (compared with our work and Dieuleveut and Bach (2015)): they require the norm $\|\mathbf{H}^{-\alpha}\mathbf{w}^*\|_2$ to be bounded for $\alpha \geq 0$ (see Theorem 1 condition (a) in Berthier et al. (2020)). In contrast, as we have discussed after Theorem 2.1, our work and Dieuleveut and Bach (2015) only require the source condition $\|\mathbf{H}^{-\alpha}\mathbf{w}^*\|_2 < \infty$ hold for $\alpha \geq -1/2$. Lastly, we would like to point out that our fourth moment assumption (Assumption 2.2) is a natural starting point for analyzing the over-parameterized regime because it also allows for direct comparisons to OLS and ridge regression, as discussed above.

Another series of prior works attempt to explain the generalization ability of SGD through the lens of implicit regularization. For overparameterized least square problems, it has been shown that multi-pass SGD converges to the minimum-norm solution (Neyshabur et al., 2015; Gunasekar et al., 2018). However, it is arguable that such a norm-based implicit regularization may not be sufficient to explain the generalization of SGD (Dauber et al., 2020; Sekhari et al., 2021; Amir et al., 2021) for general convex optimization problems. It has been shown that for some stochastic convex optimization problems, there is a sample complexity separation between SGD and the regularized empirical risk minimization (RERM) (Dauber et al., 2020; Sekhari et al., 2021) or the batch GD (Amir et al., 2021) (a followup work by Amir et al. (2022) showed that GD can match the sample complexity of SGD when restricted to stochastic convex optimization problems in the generalized linear form). Our work can be seen as an extension of existing implicit regularization works for least square problems, which builds a connection between the single-pass SGD and the ridge regression solution in an instance-wise manner. As we have claimed in Section 2.2, SGD with iterate averaging and constant learning rate is comparable to ridge regression with a constant regularization parameter, which partially explains the implicit regularization effect of SGD for least square problems. One of the authors' follow-up works (Zou et al., 2021) has extended the analysis on the implicit regularization of SGD by directly comparing the required sample sizes of SGD and ridge regression when the learning rate of SGD and the regularization parameter of ridge regression are allowed to be tuned, which precisely characterizes the least square problem instances that are preferable for SGD or ridge regression. Besides, it is also interesting to explore whether an instance-wise comparison between SGD and RERM can be obtained for other convex optimization problems (e.g., logistic regression).

Concurrent to this work, Chen et al. (2020) provide dimension independent bounds for averaged SGD; their excess risk bounds for linear regression are not as sharp as those provided here.

4. Proof Outline

We now provide the high level ideas in the proof. A key idea is relating the finite sample (infinite dimensional) covariance matrices of the variance and bias stochastic processes to those of their corresponding (asymptotic) stationary covariance matrices — an idea developed in Jain et al. (2017a) for the finite dimensional, variance analysis.

This section is organized as follows: Section 4.1 introduces additional notation and relevant linear operators; Section 4.2 presents a refined bound on a now standard bias-variance decomposition; Section 4.3 outlines the variance error analysis, followed by Section 4.4 outlining the bias error analysis. Complete proofs of the upper and lower bounds are provided in the Appendix B and Appendix C, respectively.

4.1 Preliminaries

For two matrices $\mathbf{A}$ and $\mathbf{B}$, their inner product is defined as $\langle \mathbf{A}, \mathbf{B}\rangle := \operatorname{tr}\big(\mathbf{A}^\top \mathbf{B}\big)$. The following properties will be used frequently: if $\mathbf{A}$ is PSD, and $\mathbf{B} \succeq \mathbf{B}'$, then $\langle \mathbf{A}, \mathbf{B}\rangle \geq \langle \mathbf{A}, \mathbf{B}'\rangle$. We use $\otimes$ to denote the kronecker/tensor product. We define the following linear operators:

$$\mathcal{I} = \mathbf{I}\otimes\mathbf{I}, \quad \mathcal{M} = \mathbb{E}[\mathbf{x}\otimes\mathbf{x}\otimes\mathbf{x}\otimes\mathbf{x}], \quad \widetilde{\mathcal{M}} = \mathbf{H}\otimes\mathbf{H},$$
$$\mathcal{T} = \mathbf{H}\otimes\mathbf{I} + \mathbf{I}\otimes\mathbf{H} - \gamma\mathcal{M}, \quad \widetilde{\mathcal{T}} = \mathbf{H}\otimes\mathbf{I} + \mathbf{I}\otimes\mathbf{H} - \gamma\mathbf{H}\otimes\mathbf{H}.$$

We use the notation $\mathcal{O}\circ\mathbf{A}$ to denotes the operator $\mathcal{O}$ acting on a symmetric matrix $\mathbf{A}$. For example, with these definitions, we have that for a symmetric matrix $\mathbf{A}$,

$$\mathcal{I}\circ\mathbf{A} = \mathbf{A}, \quad \mathcal{M}\circ\mathbf{A} = \mathbb{E}[(\mathbf{x}^\top\mathbf{A}\mathbf{x})\mathbf{x}\mathbf{x}^\top], \quad \widetilde{\mathcal{M}}\circ\mathbf{A} = \mathbf{H}\mathbf{A}\mathbf{H},$$
$$(\mathcal{I}-\gamma\mathcal{T})\circ\mathbf{A} = \mathbb{E}[(\mathbf{I}-\gamma\mathbf{x}\mathbf{x}^\top)\mathbf{A}(\mathbf{I}-\gamma\mathbf{x}\mathbf{x}^\top)], \quad (\mathcal{I}-\gamma\widetilde{\mathcal{T}})\circ\mathbf{A} = (\mathbf{I}-\gamma\mathbf{H})\mathbf{A}(\mathbf{I}-\gamma\mathbf{H}). \tag{4.1}$$

We conclude by summarizing a few technical properties of these operators (see Lemma B.1 in Appendix).

Lemma 4.1 *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 $0<\gamma\leq 1/\lambda_1$, then $\widetilde{\mathcal{T}}^{-1}$ exists, and is a PSD mapping.*
5. *If $0<\gamma\leq 1/(\alpha\operatorname{tr}(\mathbf{H}))$, then $\mathcal{T}^{-1}\circ\mathbf{A}$ exists for PSD matrix $\mathbf{A}$, and $\mathcal{T}^{-1}$ is a PSD mapping.*

4.2 The Bias-Variance Decomposition

It is helpful to consider the bias-variance decomposition for averaged SGD, which has been extensively studied before in the underparameterized regime ($N\gg d$) (Dieuleveut and Bach, 2015; Jain et al., 2017b,a). For convenience, we define the centered SGD iterate as $\boldsymbol{\eta}_t := \mathbf{w}_t - \mathbf{w}^*$. Similarly we define $\bar{\boldsymbol{\eta}}_N := \frac{1}{N}\sum_{t=0}^{N-1}\boldsymbol{\eta}_t$.
(1) If the sampled data contains no label noise, i.e., $y_t = \langle\mathbf{w}^*, \mathbf{x}_t\rangle$, then the obtained SGD iterates $\{\boldsymbol{\eta}_t^{\mathrm{bias}}\}$ reveal the *bias error*,

$$\boldsymbol{\eta}_t^{\mathrm{bias}} = \Big(\mathbf{I}-\gamma\mathbf{x}_t\mathbf{x}_t^\top\Big)\boldsymbol{\eta}_{t-1}^{\mathrm{bias}}, \qquad \boldsymbol{\eta}_0^{\mathrm{bias}} = \boldsymbol{\eta}_0. \tag{4.2}$$

(2) If the iterates are initialized from the optimal $\mathbf{w}^*$, i.e., $\mathbf{w}_0 = \mathbf{w}^*$, then the obtained SGD iterates $\{\boldsymbol{\eta}_t^{\mathrm{variance}}\}$ reveal the *variance error*,

$$\boldsymbol{\eta}_t^{\mathrm{variance}} = \Big(\mathbf{I}-\gamma\mathbf{x}_t\mathbf{x}_t^\top\Big)\boldsymbol{\eta}_{t-1}^{\mathrm{variance}} + \gamma\xi_t\mathbf{x}_t, \qquad \boldsymbol{\eta}_0^{\mathrm{variance}} = \mathbf{0}, \tag{4.3}$$

where $\xi_t := y_t - \langle \mathbf{w}^*, \mathbf{x}_t \rangle$ is the inherent noise. Note the "bias iterates" can be viewed as a stochastic process of SGD on a consistent linear system; similarly, the "variance iterates" should be treated as a stochastic process of SGD initialized from the optimum.

Using the defined operators, the update rule of the iterates (4.2) imply the following recursive form of $\mathbf{B}_t := \mathbb{E}[\boldsymbol{\eta}_t^{\mathrm{bias}} \otimes \boldsymbol{\eta}_t^{\mathrm{bias}}]$:

$$\mathbf{B}_t = (\mathcal{I} - \gamma \mathcal{T}) \circ \mathbf{B}_{t-1} \qquad \text{and} \qquad \mathbf{B}_0 = \boldsymbol{\eta}_0 \otimes \boldsymbol{\eta}_0, \tag{4.4}$$

and the update rule (4.3) imply the following recursive form of $\mathbf{C}_t := \mathbb{E}[\boldsymbol{\eta}_t^{\mathrm{variance}} \otimes \boldsymbol{\eta}_t^{\mathrm{variance}}]$:

$$\mathbf{C}_t = (\mathcal{I} - \gamma \mathcal{T}) \circ \mathbf{C}_{t-1} + \gamma^2 \boldsymbol{\Sigma}, \qquad \mathbf{C}_0 = \mathbf{0}. \tag{4.5}$$

We define the averaged version of $\boldsymbol{\eta}_t^{\mathrm{bias}}$ and $\boldsymbol{\eta}_t^{\mathrm{variance}}$ in the same way as $\overline{\mathbf{w}}_N$, i.e., $\bar{\boldsymbol{\eta}}_N^{\mathrm{bias}} := \frac{1}{N}\sum_{t=0}^{N-1} \boldsymbol{\eta}_t^{\mathrm{bias}}$ and $\bar{\boldsymbol{\eta}}_N^{\mathrm{variance}} := \frac{1}{N}\sum_{t=0}^{N-1} \boldsymbol{\eta}_t^{\mathrm{variance}}$. With a little abuse of probability space, from (1.2), (4.2) and (4.3) we have that

$$\boldsymbol{\eta}_t = \boldsymbol{\eta}_t^{\mathrm{bias}} + \boldsymbol{\eta}_t^{\mathrm{variance}},$$

then an application of Cauchy–Schwarz inequality leads to the following *bias-variance decomposition* on the excess risk (see Jain et al. (2017b), also Lemma B.2 in the appendix):

$$\mathbb{E}[L(\overline{\mathbf{w}}_N)] - L(\mathbf{w}^*) = \frac{1}{2}\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\eta}}_N \otimes \bar{\boldsymbol{\eta}}_N] \rangle \le \left(\sqrt{\mathrm{bias}} + \sqrt{\mathrm{variance}}\right)^2, \tag{4.6}$$

$$\text{where} \quad \mathrm{bias} := \frac{1}{2}\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\eta}}_N^{\mathrm{bias}} \otimes \bar{\boldsymbol{\eta}}_N^{\mathrm{bias}}] \rangle, \quad \mathrm{variance} := \frac{1}{2}\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\eta}}_N^{\mathrm{variance}} \otimes \bar{\boldsymbol{\eta}}_N^{\mathrm{variance}}] \rangle.$$

In the above bound, the two terms are usually referred to as the *bias error* and the *variance error* respectively. Furthermore, expanding the kronecker product between the two averaged iterates, and doubling the squared terms, we have the following upper bounds on the bias error and the variance error (see Lemma B.3 in the appendix for the proof):

$$\mathrm{bias} := \frac{1}{2}\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\eta}}_N^{\mathrm{bias}} \otimes \bar{\boldsymbol{\eta}}_N^{\mathrm{bias}}] \rangle \le \frac{1}{N^2} \sum_{t=0}^{N-1} \sum_{k=t}^{N-1} \big\langle (\mathbf{I} - \gamma \mathbf{H})^{k-t} \mathbf{H}, \mathbf{B}_t \big\rangle, \tag{4.7}$$

$$\mathrm{variance} := \frac{1}{2}\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\eta}}_N^{\mathrm{variance}} \otimes \bar{\boldsymbol{\eta}}_N^{\mathrm{variance}}] \le \frac{1}{N^2} \sum_{t=0}^{N-1} \sum_{k=t}^{N-1} \big\langle (\mathbf{I} - \gamma \mathbf{H})^{k-t} \mathbf{H}, \mathbf{C}_t \big\rangle. \tag{4.8}$$

Note that in the above bounds, we keep both summations in finite steps, and this makes our analysis sharp as $N \ll d$. In comparison, Jain et al. (2017a,b) take the inner summation to infinity, which yields looser upper bounds for further analysis in the overparameterized setting. Next we bound the two error terms (4.7) and (4.8) separately.

4.3 Bounding the Variance Error

We would like to point out that in the analysis of the variance error (4.8), Assumption 2.2 can be replaced by a weaker assumption: $\mathbb{E}[\mathbf{x}\mathbf{x}^\top \mathbf{x}\mathbf{x}^\top] \preceq R^2 \mathbf{H}$, where R is a positive constant (Jain et al., 2017b,a; Dieuleveut et al., 2017). A proof under the weaker assumption can be found in Appendix B.3. Here, for consistency, we sketch the proof under Assumption 2.2.

To upper bound (4.8), noticing that $(\mathbf{I}-\gamma\mathbf{H})^{k-t}\mathbf{H}$ is PSD, it suffices to upper bound $\mathbf{C}_t$ in PSD sense. In particular, by Lemma 5 in Jain et al. (2017a) (restated in Lemma B.4 in the appendix), the sequence $\{\mathbf{C}_t\}_{t=0,\ldots}$ has the following property,

$$0 = \mathbf{C}_0 \preceq \mathbf{C}_1 \preceq \cdots \preceq \mathbf{C}_\infty \preceq \frac{\gamma\sigma^2}{1-\gamma\alpha\,\mathrm{tr}(\mathbf{H})}\mathbf{I}. \tag{4.9}$$

This gives a uniform but crude upper bound on $\mathbf{C}_t$ for all $t \ge 0$. However, a direct application of this crude bound to (4.8) cannot give a sharp rate in the overparameterized setting. Instead, we seek to refine the bound of $\mathbf{C}_t$ based on its update rule in (4.5) (see the proof of Lemma B.5 for details):

$$\begin{aligned}
\mathbf{C}_t &= (\mathcal{I}-\gamma\mathcal{T})\circ\mathbf{C}_{t-1} + \gamma^2\mathbf{\Sigma}\\
&= (\mathcal{I}-\gamma\widetilde{\mathcal{T}})\circ\mathbf{C}_{t-1} + \gamma^2(\mathcal{M}-\widetilde{\mathcal{M}})\circ\mathbf{C}_{t-1} + \gamma^2\mathbf{\Sigma}\\
&\preceq (\mathcal{I}-\gamma\widetilde{\mathcal{T}})\circ\mathbf{C}_{t-1} + \gamma^2\mathcal{M}\circ\mathbf{C}_{t-1} + \gamma^2\mathbf{\Sigma} \qquad (\text{since } \widetilde{\mathcal{M}} \text{ is a PSD mapping})\\
&\preceq (\mathcal{I}-\gamma\widetilde{\mathcal{T}})\circ\mathbf{C}_{t-1} + \frac{\gamma^3\sigma^2}{1-\gamma\alpha\,\mathrm{tr}(\mathbf{H})}\mathcal{M}\circ\mathbf{I} + \gamma^2\mathbf{\Sigma}, \quad (\text{by (4.9) and } \mathcal{M} \text{ is a PSD mapping})\\
&\preceq (\mathcal{I}-\gamma\widetilde{\mathcal{T}})\circ\mathbf{C}_{t-1} + \frac{\gamma^3\sigma^2\alpha\,\mathrm{tr}(\mathbf{H})}{1-\gamma\alpha\,\mathrm{tr}(\mathbf{H})}\mathbf{H} + \gamma^2\sigma^2\mathbf{H}, \qquad (\text{by Assumptions 2.2 and 2.3})\\
&= (\mathcal{I}-\gamma\widetilde{\mathcal{T}})\circ\mathbf{C}_{t-1} + \frac{\gamma^2\sigma^2}{1-\gamma\alpha\,\mathrm{tr}(\mathbf{H})}\mathbf{H}.
\end{aligned}$$

Solving the above recursion, we obtain the following refined upper bound for $\mathbf{C}_t$:

$$\begin{aligned}
\mathbf{C}_t &\preceq \frac{\gamma^2\sigma^2}{1-\gamma\alpha\,\mathrm{tr}(\mathbf{H})}\sum_{k=0}^{t-1}(\mathcal{I}-\gamma\widetilde{\mathcal{T}})^k\circ\mathbf{H}\\
&= \frac{\gamma^2\sigma^2}{1-\gamma\alpha\,\mathrm{tr}(\mathbf{H})}\sum_{k=0}^{t-1}(\mathbf{I}-\gamma\mathbf{H})^k\mathbf{H}(\mathbf{I}-\gamma\mathbf{H})^k \qquad (\text{by the property of } \mathcal{I}-\gamma\widetilde{\mathcal{T}} \text{ in (4.1)})\\
&\preceq \frac{\gamma^2\sigma^2}{1-\gamma\alpha\,\mathrm{tr}(\mathbf{H})}\sum_{k=0}^{t-1}(\mathbf{I}-\gamma\mathbf{H})^k\mathbf{H} = \frac{\gamma\sigma^2}{1-\gamma\alpha\,\mathrm{tr}(\mathbf{H})}\cdot\big(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^t\big).
\end{aligned} \tag{4.10}$$

Now we can plug the above refined upper bound (4.10) into (4.8), and obtain

$$\begin{aligned}
\text{variance} &\le \frac{\sigma^2}{N^2(1-\gamma\alpha\,\mathrm{tr}(\mathbf{H}))}\sum_{t=0}^{N-1}\big\langle \mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N-t}, \mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^t\big\rangle\\
&= \frac{\sigma^2}{N^2(1-\gamma\alpha\,\mathrm{tr}(\mathbf{H}))}\sum_{t=0}^{N-1}\sum_i\big(1-(1-\gamma\lambda_i)^{N-t}\big)\big(1-(1-\gamma\lambda_i)^t\big)\\
&\le \frac{\sigma^2}{N^2(1-\gamma\alpha\,\mathrm{tr}(\mathbf{H}))}\cdot N\cdot\sum_i\big(1-(1-\gamma\lambda_i)^N\big)^2.
\end{aligned} \tag{4.11}$$

The remaining effort is to precisely control the summations in (4.11) according to the scale of the eigenvalues: for large eigenvalues $\lambda_i \ge \frac{1}{N\gamma}$, which appear at most k^* times, we use $1-(1-\gamma\lambda_i)^N \le 1$; and for the remaining small eigenvalues $\lambda_i < \frac{1}{N\gamma}$, we use $1-(1-\gamma\lambda_i)^N \le \mathcal{O}\left(N\gamma\lambda_i\right)$. Plugging these into (4.11) gives us the final full spectrum upper bound on the variance error (see the proof of Lemma B.6 for more details). This bound contributes to part of EffectiveVar in Theorem 2.1.

4.4 Bounding the Bias Error

Next we discuss how to bound the bias error (4.7). A natural idea is to follow the same way in analyzing the variance error, and derive a similar bound on $\mathbf{B}_t$. Yet a fundamental difference between the variance sequence (4.5) and the bias sequence (4.4) is that: $\mathbf{C}_t$ is increasing, while $\mathbf{B}_t$ is "contracting", hence applying the same procedure in the variance error analysis cannot lead to a tight bound on $\mathbf{B}_t$. Instead, observing that $\mathbf{S}_t := \sum_{k=0}^{t-1}\mathbf{B}_k$, the summation of a contracting sequence, is increasing in the PSD sense. Particularly, we can rewrite $\mathbf{S}_t$ in the following recursive form

$$\mathbf{S}_t = (\mathcal{I}-\gamma\mathcal{T})\circ\mathbf{S}_{t-1} + \mathbf{B}_0, \tag{4.12}$$

which resembles that of $\mathbf{C}_t$ in (4.5). This motivates us to: (i) express the obtained bias error bound (4.7) by $\mathbf{S}_t$, and (ii) derive a tight upper bound on $\mathbf{S}_t$ using similar analysis for the variance error.

For (i), by some linear algebra manipulation (see the derivation of (B.13)), we can bound (4.7) as follows:

$$\mathrm{bias} \le \frac{1}{\gamma N^2}\Big\langle \mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^N, \sum_{t=0}^{N-1}\mathbf{B}_t\Big\rangle = \frac{1}{\gamma N^2}\big\langle \mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^N, \mathbf{S}_N\big\rangle. \tag{4.13}$$

For (ii), we first show that $\{\mathbf{S}_t\}_{t=1,\dots,N}$ is increasing and has a crude upper bound (see Lemmas B.7 and B.9):

$$\mathbf{B}_0 = \mathbf{S}_1 \preceq \mathbf{S}_2 \preceq \cdots \preceq \mathbf{S}_N, \quad \text{and} \quad \mathcal{M}\circ\mathbf{S}_N \preceq \frac{\alpha\cdot\mathrm{tr}\Big(\big(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{2N}\big)\mathbf{B}_0\Big)}{\gamma(1-\gamma\alpha\,\mathrm{tr}(\mathbf{H}))}\cdot\mathbf{H}. \tag{4.14}$$

Then similar to our previous procedure in bounding $\mathbf{C}_t$, we can tighten the upper bound on $\mathbf{S}_t$ by its recursive form (4.12) and the crude bound ($\mathcal{M}\circ\mathbf{S}_{N-1}$ in (4.14)), and obtain the following refined bound (see Lemma B.10) for $\mathbf{S}_N$:

$$\mathbf{S}_N \preceq \sum_{k=0}^{N-1}(\mathbf{I}-\gamma\mathbf{H})^k\mathbf{B}_0(\mathbf{I}-\gamma\mathbf{H})^k + \frac{\gamma\alpha\cdot\mathrm{tr}\Big(\big(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{2N}\big)\mathbf{B}_0\Big)}{1-\gamma\alpha\,\mathrm{tr}(\mathbf{H})}\sum_{k=0}^{N-1}(\mathbf{I}-\gamma\mathbf{H})^{2k}\mathbf{H}. \tag{4.15}$$

The remaining proof will be similar to what we have done for the variance error bound: substituting (4.15) into (4.13) gives an upper bound on the bias error with respect to the summations over functions of eigenvalues. Then by carefully controlling each summation according to the scale of the corresponding eigenvalues, we will obtain a tight full spectrum upper bound on the bias error (see the proof of Lemma B.11 for more details).

As a final remark, noticing that different from the upper bound of $\mathbf{C}_t$ in (4.10), the upper bound for $\mathbf{S}_t$ in (4.15) consists of two terms. The first term will contribute to the EffectiveBias term in Theorem 2.1, while the second term will be merged to the bound of the variance error and contribute to the EffectiveVar term in Theorem 2.1.

5. The Effect of Tail-Averaging

We further consider benign overfitting of SGD when *tail-averaging* (Polyak and Juditsky, 1992; Jain et al., 2017a,b) is applied, i.e.,

$$\overline{\mathbf{w}}_{s:s+N} = \frac{1}{N}\sum_{t=s}^{s+N-1}\mathbf{w}_t.$$

We present the following theorem as a counterpart of Theorem 2.1. The proof is deferred to Appendix D.

Theorem 5.1 (Benign overfitting of SGD with tail-averaging) *Consider SGD with tail-averaging. Suppose Assumptions 2.1-2.3 hold and that the stepsize is set so that $\gamma < 1/(\alpha\,\mathrm{tr}(\mathbf{H}))$. Then the excess risk can be upper bounded as follows,*

$$\mathbb{E}[L(\overline{\mathbf{w}}_{s:s+N})] - L(\mathbf{w}^*) \le 2\cdot \mathrm{EffectiveBias} + 2\cdot\mathrm{EffectiveVar},$$

where

$$\begin{aligned}\mathrm{EffectiveBias} &= \frac{1}{\gamma^2N^2}\cdot\big\|(\mathbf{I}-\gamma\mathbf{H})^s(\mathbf{w}_0-\mathbf{w}^*)\big\|^2_{\mathbf{H}^{-1}_{0:k^*}} + \big\|(\mathbf{I}-\gamma\mathbf{H})^s(\mathbf{w}_0-\mathbf{w}^*)\big\|^2_{\mathbf{H}_{k^*:\infty}}\\ \mathrm{EffectiveVar} &= \frac{4\alpha\big(\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{I}_{0:k^\dagger}} + (s+N)\gamma\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{H}_{k^\dagger:\infty}}\big)}{N\gamma(1-\gamma\alpha\,\mathrm{tr}(\mathbf{H}))}\cdot\bigg(\frac{k^*}{N}+N\gamma^2\sum_{i>k^*}\lambda_i^2\bigg)\\ &\quad+\frac{\sigma^2}{1-\gamma\alpha\,\mathrm{tr}(\mathbf{H})}\cdot\bigg(\frac{k^*}{N}+\gamma\cdot\sum_{k^*<i\le k^\dagger}\lambda_i+(s+N)\gamma^2\cdot\sum_{i>k^\dagger}\lambda_i^2\bigg),\end{aligned}$$

where $k^* = \max\{k:\lambda_k\ge\frac{1}{\gamma N}\}$ and $k^\dagger=\max\{k:\lambda_k\ge\frac{1}{\gamma(s+N)}\}$.

Theorem 5.1 shows that tail-averaging has improvements over iterate-averaging. This agrees with the results shown in Jain et al. (2017b): in the underparameterized regime ($N \gg d$) and for the strongly convex case ($\lambda_d > 0$), one can obtain substantially improved convergence rates on the bias term.

We also provide a lower bound on the excess risk for SGD with tail-averaging as a counterpart of Theorem 2.2, which shows that our upper bound is nearly tight. The proof is again deferred to Appendix D.

Theorem 5.2 (Excess risk lower bound, tail-averaging) *Consider SGD with tail-averaging. Suppose $N\ge 500$. For any well-specified data distribution $\mathcal{D}$ (see (2.1)) that also satisfies Assumptions 2.1, 2.2 and 2.4, for any stepsize such that $\gamma<1/\lambda_1$, we have that:*

$$\begin{aligned}\mathbb{E}[L(\overline{\mathbf{w}}_{s:s+N})]-L(\mathbf{w}^*) &\ge \frac{1}{100\gamma^2N^2}\cdot\|(\mathbf{I}-\gamma\mathbf{H})^s(\mathbf{w}_0-\mathbf{w}^*)\|^2_{\mathbf{H}^{-1}_{0:k^*}}+\frac{\|(\mathbf{I}-\gamma\mathbf{H})^s(\mathbf{w}_0-\mathbf{w}^*)\|^2_{\mathbf{H}_{k^*:\infty}}}{100}\\ &\quad+\frac{\beta\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{H}_{k^\dagger:\infty}}}{16000}\bigg(\frac{k^*}{N}+N\gamma^2\sum_{i>k^*}\lambda_i^2\bigg)\end{aligned}$$

$$+\frac{\sigma_{\text{noise}}^2}{600}\left(\frac{k^*}{N}+\gamma\sum_{k^*<i\le k^\dagger}\lambda_i+(s+N)\gamma^2\sum_{i>k^\dagger}\lambda_i^2\right),$$

where $k^*=\max\{k:\lambda_k\ge\frac{1}{N\gamma}\}$ *and* $k^\dagger=\max\{k:\lambda_k\ge\frac{1}{(s+N)\gamma}\}$.

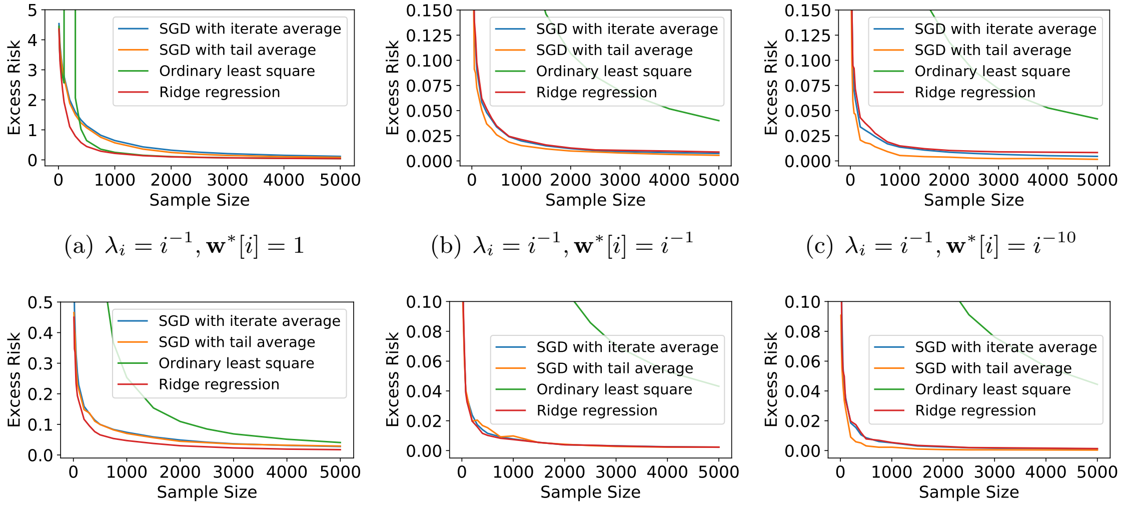
Comparing our upper and lower bounds, they are matching (upto absolute constants) for most of the terms, except for the first effective variance term, where a $\|\mathbf{w}_0-\mathbf{w}^*\|_{\mathbf{I}_{0:k^\dagger}}^2$ is lost (suppose that $s=\Theta(N)$). Our conjecture is that the upper bound is improvable in this regard. Obtaining matching upper and lower bounds for SGD with tail-averaging is left as a direction for future work.

6. Experiments

In this section, we seek to empirically observe the benign overfitting phenomenon for SGD in Gaussian least square problems and verify our theorems on the generalization performance of SGD.

We first consider three over-parameterized linear regression problem instances with $d=2000$ and the spectrum of $\mathbf{H}$ as $\lambda_i=i^{-1}$, $\lambda_i=i^{-1}\log(i)^{-2}$, and $\lambda_i=i^{-2}$, respectively. Besides, the ground truth is fixed to be $\mathbf{w}^*[i]=i^{-1}$. The training and test risks for these three problems are displayed in Figure 1. We observe that when $\lambda_i=i^{-1}$, the SGD algorithm overfits the training data and fails to generalize; when $\lambda_i=i^{-1}\log(i)^{-2}$, SGD overfits the training data (achieving a training risk much smaller than the Bayes risk) while generalizes well (achieving a vanishing test risk), which exhibits a benign overfitting phenomenon of SGD; when $\lambda_i=i^{-2}$, SGD gives vanishing test risk and tends to un-fit the training data, which indicates a regularization effect of SGD. In sum, the experiments suggest that benign overfitting of SGD can happen when the spectrum of $\mathbf{H}$ decays neither fast nor slow. This is consistent with the benign overfitting of least square (minimum-norm solution) (Bartlett et al., 2020), where for $\mathbf{H}$ with spectrum in form of $\lambda_i=i^{-\alpha}\log^{-\beta}(i)$, the benign overfitting phenomenon can only happen for $\alpha=1$ and $\beta>1$.

Then we consider 6 problem instances, which are the combinations of two covariance matrices $\mathbf{H}$ with eigenvalues $\lambda_i=i^{-1}$ and $\lambda_i=i^{-2}$ respectively; and three true model parameter $\mathbf{w}^*$ with components $\mathbf{w}^*[i]=1$, $\mathbf{w}^*[i]=i^{-1}$, and $\mathbf{w}^*[i]=i^{-10}$, respectively. We investigate four algorithms: (1) SGD with iterate averaging (from the beginning), (2) SGD with tail averaging ($\bar{\mathbf{w}}_{N/2:N-1}$), (3) ordinary least square (minimum-norm interpolator), and (4) ridge regression (regularized least square), where the hyperparameters (i.e., γ for SGD and λ for ridge regression) are fine-tuned to achieve the best performance. Results are shown in Figure 2. We see that (1) SGD, with either iterate averaging or tail averaging, is comparable to ridge regression, and significantly outperforms ordinary least square in some problem instances, and (2) SGD with tail averaging performs better than SGD with iterate averaging. These observations are consistent with our theoretical findings and demonstrate the benefit of the implicit regularization from SGD.



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

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

(c) $\lambda_i = i^{-1}, \mathbf{w}^*[i] = i^{-10}$

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

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

(f) $\lambda_i = i^{-2}, \mathbf{w}^*[i] = i^{-10}$

Figure 2: Excess risk comparison between SGD with iterate average, SGD with tail average, ordinary least square, and ridge regression, where the stepsize γ and regularization parameter λ are fine-tuned to achieve the best performance. The problem dimension is $d = 200$ and the variance of model noise is $\sigma^2 = 1$. We consider 6 combinations of 2 different covariance matrices and 3 different ground truth model vectors. The plots are averaged over 20 independent runs.

7. Discussion

This work considers the question of how well constant-stepsize SGD (with iterate average or tail average) generalizes for the linear regression problem in the overparameterized regime. Our main result provides a sharp excess risk bound, stated in terms of the full eigenspectrum of the data covariance matrix. Our results reveal how a benign-overfitting phenomenon can occur under certain spectrum decay conditions on the data covariance.

There are number of more subtle points worth reflecting on:

Moving beyond the square loss. Focusing on linear regression is a means to understand phenomena that are exhibited more broadly. One natural next step here would be understand the analogues of the classical iterate averaging results (Polyak and Juditsky, 1992) for locally quadratic models, where decaying stepsizes are necessary for vanishing risk.

Relaxing the data distribution assumption. While our data distribution assumption (Assumption 2.2) can be satisfied if the whitened data is sub-Gaussian, it still cannot cover the simple one-hot case (i.e., $\mathbf{x} = \mathbf{e}_i$ with probability p_i, where $\sum_i p_i = 1$). Here, we conjecture that modifications of our proof can be used to establish the theoretical guarantees of SGD under the following relaxed assumption on the data distribution: assume that $\mathbb{E}[\mathbf{x}\mathbf{x}^\top \mathbf{A}\mathbf{x}\mathbf{x}^\top] \le a \operatorname{tr}(\mathbf{H}\mathbf{A}) \cdot \mathbf{H} + b\|\mathbf{H}\|_2 \cdot \mathbf{H}^{1/2}\mathbf{A}\mathbf{H}^{1/2}$ for all PSD matrix $\mathbf{A}$ and some nonnegative constants a and b, which is weaker than Assumption 2.2 in the sense that we can allow $a = 0$; this assumption captures the case where $\mathbf{x}$ are standard basis vectors, with $a = 0$ and $b = 1$.

Acknowledgement

Difan Zou was supported by the Bloomberg Data Science Ph.D. Fellowship. Jingfeng Wu was supported in part by NSF CAREER grant 1652257. Vladimir Braverman is partially supported by the National Science Foundation Awards 2244899 and 2333887 and the ONR award N000142312737. Quanquan Gu is supported in part by the National Science Foundation CAREER Award 1906169 and IIS-2008981. Sham Kakade acknowledges funding from the Office of Naval Research under award N00014-22-1-2377 and the National Science Foundation Grant under award CCF-2212841.

References

Idan Amir, Tomer Koren, and Roi Livni. Sgd generalizes better than gd (and regularization doesn't help). In *Conference on Learning Theory*, pages 63–92. PMLR, 2021.

Idan Amir, Roi Livni, and Nathan Srebro. Thinking outside the ball: Optimal learning with gradient descent for generalized linear stochastic convex optimization. *arXiv preprint arXiv:2202.13328*, 2022.

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

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

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

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

Niladri S Chatterji and Philip M Long. Finite-sample analysis of interpolating linear classifiers in the overparameterized regime. *arXiv preprint arXiv:2004.12019*, 2020.

Xi Chen, Qiang Liu, and Xin T Tong. Dimension independent generalization error with regularized online optimization. *arXiv preprint arXiv:2003.11196*, 2020.

Assaf Dauber, Meir Feder, Tomer Koren, and Roi Livni. Can implicit bias explain generalization? stochastic convex optimization as a case study. *Advances in Neural Information Processing Systems*, 33:7743–7753, 2020.

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

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

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

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

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

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

Prateek Jain, Sham M Kakade, Rahul Kidambi, Praneeth Netrapalli, Venkata Krishna Pillutla, and Aaron Sidford. A markov chain theory approach to characterizing the minimax optimality of stochastic gradient descent (for least squares). *arXiv preprint arXiv:1710.09430*, 2017a.

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

Prateek Jain, Sham M. Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Accelerating stochastic gradient descent for least squares regression. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*. PMLR, 2018.

Chandrashekar Lakshminarayanan and Csaba Szepesvari. Linear stochastic approximation: How far does constant step-size and iterate averaging go? In *International Conference on Artificial Intelligence and Statistics*, pages 1347–1355, 2018.

Vidya Muthukumar, Adhyyan Narang, Vignesh Subramanian, Mikhail Belkin, Daniel Hsu, and Anant Sahai. Classification vs regression in overparameterized regimes: Does the loss function matter? *arXiv preprint arXiv:2005.08054*, 2020.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *arXiv preprint arXiv:1912.02292*, 2019.

Preetum Nakkiran, Prayaag Venkat, Sham Kakade, and Tengyu Ma. Optimal regularization can mitigate double descent. *arXiv preprint arXiv:2003.01897*, 2020.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. In *ICLR (Workshop)*, 2015.

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

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *ICML*, 2012.

Bernhard Schölkopf, Alexander J Smola, Francis Bach, et al. *Learning with kernels: support vector machines, regularization, optimization, and beyond.* MIT press, 2002.

Ayush Sekhari, Karthik Sridharan, and Satyen Kale. Sgd: The role of implicit regularization, batch-size and multiple-epochs. *Advances In Neural Information Processing Systems*, 34: 27422–27433, 2021.

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

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

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

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

Appendix A. Discussions on Assumption 2.2

Hsu et al. (2014); Bartlett et al. (2020); Tsigler and Bartlett (2020) assume that $\mathbf{z} := \mathbf{H}^{-\frac{1}{2}}\mathbf{x}$ is sub-Gaussian. The following lemma shows that our Assumption 2.2 is implied by assuming sub-Gaussianity.

Lemma A.1 *Suppose* $\mathbb{E}[\mathbf{x}\mathbf{x}^\top] = \mathbf{H}$*, and* $\mathbf{z} := \mathbf{H}^{-\frac{1}{2}}\mathbf{x}$ *is* σ_z^2*-sub-Gaussian random vector, then for any PSD matrix* $\mathbf{A}$*, we have*

$$\mathbb{E}[(\mathbf{x}^\top\mathbf{A}\mathbf{x})\mathbf{x}\mathbf{x}^\top] \preceq 16\sigma_z^4 \operatorname{tr}(\mathbf{A}\mathbf{H})\mathbf{H}.$$

Proof Note that $\mathbf{z}$ is a σ_z^2-sub-Gaussian random vector with identity covariance matrix, implying that for any fixed unit vector $\mathbf{u}$ that $\mathbf{u}^\top\mathbf{z}$ is a σ_z^2-sub-Gaussian random variable. Then we have the following inequality for any unit vectors $\mathbf{u}$ and $\mathbf{v}$

$$\mathbb{E}[(\mathbf{u}^\top\mathbf{z})^2(\mathbf{v}^\top\mathbf{z})^2] \le \sqrt{\mathbb{E}[(\mathbf{u}^\top\mathbf{z})^4]}\cdot\sqrt{\mathbb{E}[(\mathbf{v}^\top\mathbf{z})^4]} \le \max\left\{\mathbb{E}[(\mathbf{u}^\top\mathbf{z})^4], \mathbb{E}[(\mathbf{v}^\top\mathbf{z})^4]\right\} \le 16\cdot\sigma_z^4,$$

where the first inequality follows from the Cauchy–Schwarz inequality; and the last inequality uses the fact that $\mathbf{u}^\top\mathbf{z}$ is σ_z^2 sub-Gaussian. Here, the factor 16 is due to the sub-Gaussian property (Proposition 2.5.2, Vershynin (2018)). Next, for any PSD matrix $\mathbf{A}$, suppose its eigenvalue decomposition is $\mathbf{A} = \sum_i \mu_i \mathbf{u}_i \mathbf{u}_i^\top$, where $\mu_i \ge 0$ is the eigenvalue and $\mathbf{u}_i$ is the corresponding eigenvector, we have

$$\mathbb{E}[(\mathbf{z}^\top \mathbf{A}\mathbf{z})\mathbf{z}\mathbf{z}^\top] = \sum_i \mu_i \mathbb{E}[(\mathbf{u}_i^\top \mathbf{z})^2 \mathbf{z}\mathbf{z}^\top]. \tag{A.1}$$

For any unit vector $\mathbf{v}$, we have:

$$\mathbf{v}^\top \mathbb{E}[(\mathbf{z}^\top \mathbf{A}\mathbf{z})\mathbf{z}\mathbf{z}^\top]\mathbf{v} = \sum_i \mu_i \mathbb{E}[(\mathbf{u}_i^\top \mathbf{z})^2 (\mathbf{v}^\top \mathbf{z})^2] \le 16 \cdot \sigma_z^4 \cdot \sum_i \mu_i = 16 \cdot \sigma_z^4 \operatorname{tr}(\mathbf{A}).$$

This implies that for any PSD matrix $\mathbf{A}$ we have

$$\mathbb{E}[(\mathbf{z}^\top \mathbf{A}\mathbf{z})\mathbf{z}\mathbf{z}^\top] \le 16 \cdot \sigma_z^4 \operatorname{tr}(\mathbf{A})\mathbf{I}. \tag{A.2}$$

Finally considering $\mathbf{x} = \mathbf{H}^{\frac{1}{2}}\mathbf{z}$, we have for any PSD matrix $\mathbf{A}$:

$$\begin{aligned}
\mathbb{E}[(\mathbf{x}^\top \mathbf{A}\mathbf{x})\mathbf{x}\mathbf{x}^\top] &= \mathbb{E}[(\mathbf{z}^\top \mathbf{H}^{\frac{1}{2}}\mathbf{A}\mathbf{H}^{\frac{1}{2}}\mathbf{z})\mathbf{H}^{\frac{1}{2}}\mathbf{z}\mathbf{z}^\top \mathbf{H}^{\frac{1}{2}}] \\
&= \mathbf{H}^{\frac{1}{2}}\mathbb{E}[(\mathbf{z}^\top \mathbf{H}^{\frac{1}{2}}\mathbf{A}\mathbf{H}^{\frac{1}{2}}\mathbf{z})\mathbf{z}\mathbf{z}^\top]\mathbf{H}^{\frac{1}{2}} \\
&\preceq \mathbf{H}^{\frac{1}{2}} \cdot 16\sigma_z^4 \operatorname{tr}(\mathbf{H}^{\frac{1}{2}}\mathbf{A}\mathbf{H}^{\frac{1}{2}}) \cdot \mathbf{I} \cdot \mathbf{H}^{\frac{1}{2}} \\
&= 16\sigma_z^4 \operatorname{tr}(\mathbf{A}\mathbf{H})\mathbf{H},
\end{aligned}$$

where the second line holds since $\mathbf{z}^\top \mathbf{H}^{\frac{1}{2}}\mathbf{A}\mathbf{H}^{\frac{1}{2}}\mathbf{z}$ is a scalar and the third line of the above equation is due to (A.2). This concludes the proof. ■

Appendix B. Proofs of the Upper Bounds

B.1 Technical Lemma

Lemma B.1 (Restatement of Lemma 4.1) *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 $0 < \gamma < 1/(\alpha \operatorname{tr}(\mathbf{H}))$, then $\mathcal{T}^{-1} \circ \mathbf{A}$ exists for PSD matrix $\mathbf{A}$, and $\mathcal{T}^{-1}$ is a PSD mapping.*

Proof The following proofs are summarized from Jain et al. (2017a,b), and we include them here for completeness.

1. For any PSD matrix $\mathbf{A} \succeq 0$, by definition, we have

$$\mathcal{M} \circ \mathbf{A} = \mathbb{E}[\mathbf{x}\mathbf{x}^\top \mathbf{A}\mathbf{x}\mathbf{x}^\top] \succeq 0,$$
$$\widetilde{\mathcal{M}} \circ \mathbf{A} = \mathbf{H}\mathbf{A}\mathbf{H} \succeq 0.$$

Therefore, both $\mathcal{M}$ and $\widetilde{\mathcal{M}}$ are PSD mappings.

2. For any PSD matrix $\mathbf{A} \succeq 0$, we have

$$(\mathcal{I} - \gamma\mathcal{T}) \circ \mathbf{A} = \mathbb{E}[(\mathbf{I} - \gamma\mathbf{x}\mathbf{x}^\top)\mathbf{A}(\mathbf{I} - \gamma\mathbf{x}\mathbf{x}^\top)] \succeq 0,$$
$$(\mathcal{I} - \gamma\widetilde{\mathcal{T}}) \circ \mathbf{A} = (\mathbf{I} - \gamma\mathbf{H})\mathbf{A}(\mathbf{I} - \gamma\mathbf{H}) \succeq 0.$$

Hence, $\mathcal{I} - \gamma\mathcal{T}$ and $\mathcal{I} - \gamma\widetilde{\mathcal{T}}$ are both PSD mapping.

3. For any PSD matrix $\mathbf{A} \succeq 0$,

$$(\mathcal{M} - \widetilde{\mathcal{M}}) \circ \mathbf{A} = \mathbb{E}[\mathbf{x}\mathbf{x}^\top \mathbf{A}\mathbf{x}\mathbf{x}^\top] - \mathbf{H}\mathbf{A}\mathbf{H} = \mathbb{E}[(\mathbf{x}\mathbf{x}^\top - \mathbf{H})\mathbf{A}(\mathbf{x}\mathbf{x}^\top - \mathbf{H})] \succeq 0.$$

Thus, $\widetilde{\mathcal{T}} - \mathcal{T} = \mathcal{M} - \widetilde{\mathcal{M}}$ is PSD.

4. According to (4.1), if $0 < \gamma < 1/\lambda_1$, $\mathbf{I} - \gamma\mathbf{H}$ is a contraction map, thus for any symmetric matrix $\mathbf{A}$, the following exists:

$$\sum_{t=0}^{\infty}(\mathcal{I} - \gamma\widetilde{\mathcal{T}})^t \circ \mathbf{A} = \sum_{t=0}^{\infty}(\mathbf{I} - \gamma\mathbf{H})^t\mathbf{A}(\mathbf{I} - \gamma\mathbf{H})^t.$$

Therefore, $\sum_{t=0}^{\infty}(\mathcal{I} - \gamma\widetilde{\mathcal{T}})^t$ exists and $\widetilde{\mathcal{T}}^{-1} = \gamma\sum_{t=0}^{\infty}(\mathcal{I} - \gamma\widetilde{\mathcal{T}})^t$ exists. Furthermore, for any PSD matrix $\mathbf{A} \succeq 0$, we have

$$\widetilde{\mathcal{T}}^{-1} \circ \mathbf{A} = \gamma\sum_{t=0}^{\infty}(\mathcal{I} - \gamma\widetilde{\mathcal{T}})^t \circ \mathbf{A} = \gamma\sum_{t=0}^{\infty}(\mathbf{I} - \gamma\mathbf{H})^t\mathbf{A}(\mathbf{I} - \gamma\mathbf{H})^t \succeq 0,$$

which implies $\widetilde{\mathcal{T}}^{-1}$ is a PSD mapping.

5. For any finite PSD matrix $\mathbf{A}$, consider the following identity

$$\mathcal{T}^{-1} \circ \mathbf{A} = \gamma\sum_{t=0}^{\infty}(\mathcal{I} - \gamma\mathcal{T})^t \circ \mathbf{A}.$$

Clearly, if the right hand side exists, it must be PSD since $\mathcal{I} - \gamma\mathcal{T}$ is a PSD mapping. It remains to show that $\sum_{t=0}^{\infty}(\mathcal{I} - \gamma\mathcal{T})^t \circ \mathbf{A}$ is finite, and it suffices to show that

$$\operatorname{tr}\left(\sum_{t=0}^{\infty}(\mathcal{I} - \gamma\mathcal{T})^t \circ \mathbf{A}\right) = \sum_{t=0}^{\infty}\operatorname{tr}\left((\mathcal{I} - \gamma\mathcal{T})^t \circ \mathbf{A}\right) < \infty.$$

Based on the definition of $\mathcal{T}$, let $\mathbf{A}_t = (\mathcal{I} - \gamma\mathcal{T})^t \circ \mathbf{A}$, we have

$$\operatorname{tr}(\mathbf{A}_t) = \operatorname{tr}(\mathbf{A}_{t-1}) - \gamma\operatorname{tr}(\mathbf{H}\mathbf{A}_{t-1}) - \gamma\operatorname{tr}(\mathbf{A}_{t-1}\mathbf{H}) + \gamma^2\operatorname{tr}\left(\mathbb{E}[\mathbf{x}\mathbf{x}^\top\mathbf{A}\mathbf{x}\mathbf{x}^\top]\right)$$

$$= \operatorname{tr}(\mathbf{A}_{t-1}) - 2\gamma \operatorname{tr}(\mathbf{H}\mathbf{A}_{t-1}) + \gamma^2 \operatorname{tr}\left(\mathbf{A}_{t-1}\mathbb{E}[\mathbf{x}\mathbf{x}^\top\mathbf{x}\mathbf{x}^\top]\right). \tag{B.1}$$

By Assumption 2.2, we have $\mathbb{E}[\mathbf{x}\mathbf{x}^\top\mathbf{x}\mathbf{x}^\top] \preceq \alpha \operatorname{tr}(\mathbf{H})\mathbf{H}$. Therefore, it follows that

$$\begin{aligned}\operatorname{tr}(\mathbf{A}_t) &\le \operatorname{tr}(\mathbf{A}_{t-1}) - (2\gamma - \gamma^2\alpha \operatorname{tr}(\mathbf{H}))\operatorname{tr}(\mathbf{H}\mathbf{A}_{t-1})\\ &\le \operatorname{tr}\left((\mathbf{I}-\gamma\mathbf{H})\mathbf{A}_{t-1}\right)\\ &\le (1-\gamma\lambda_d)\operatorname{tr}(\mathbf{A}_{t-1}), \end{aligned} \tag{B.2}$$

where we use the assumption $\gamma < 1/(\alpha \operatorname{tr}(\mathbf{H}))$ in the first inequality. This further implies that

$$\sum_{t=0}^{\infty} \operatorname{tr}\left((\mathcal{I}-\gamma\mathcal{T})^t \circ \mathbf{A}\right) = \sum_{t=0}^{\infty} \operatorname{tr}(\mathbf{A}_t) \le \frac{\operatorname{tr}(\mathbf{A})}{\gamma\lambda_d} < \infty.$$

Therefore, $\mathcal{T}^{-1} \circ \mathbf{A}$ exists, and is PSD. So $\mathcal{T}^{-1}$ is a PSD mapping.

■

B.2 Bias-Variance Decomposition

Lemma B.2 (Bias-variance decomposition)

$$\mathbb{E}[L(\overline{\mathbf{w}}_N)] - L(\mathbf{w}^*) = \frac{1}{2}\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\eta}}_N \otimes \bar{\boldsymbol{\eta}}_N]\rangle \le \left(\sqrt{\mathrm{bias}} + \sqrt{\mathrm{variance}}\right)^2,$$

where

$$\mathrm{bias} := \frac{1}{2}\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\eta}}_N^{\mathrm{bias}} \otimes \bar{\boldsymbol{\eta}}_N^{\mathrm{bias}}]\rangle, \qquad \mathrm{variance} := \frac{1}{2}\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\eta}}_N^{\mathrm{variance}} \otimes \bar{\boldsymbol{\eta}}_N^{\mathrm{variance}}]\rangle.$$

Proof This proof comes from (Jain et al., 2017a). For completeness we included it here.

With a slight abuse of notations (or probability spaces), we have $\boldsymbol{\eta}_t = \boldsymbol{\eta}_t^{\mathrm{bias}} + \boldsymbol{\eta}_t^{\mathrm{variance}}$, where the randomness of $\boldsymbol{\eta}_t^{\mathrm{bias}}$ and $\boldsymbol{\eta}_t^{\mathrm{variance}}$ is understood as coming from the same probability space as $\boldsymbol{\eta}_t$. This implies $\bar{\boldsymbol{\eta}}_t = \bar{\boldsymbol{\eta}}_t^{\mathrm{bias}} + \bar{\boldsymbol{\eta}}_t^{\mathrm{variance}}$. Then we have

$$\begin{aligned}&\mathbb{E}[L(\overline{\mathbf{w}}_N)] - L(\mathbf{w}^*)\\ &= \frac{1}{2}\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\eta}}_N \otimes \bar{\boldsymbol{\eta}}_N]\rangle\\ &= \mathbb{E}\left[\frac{1}{\sqrt{2}}\bar{\boldsymbol{\eta}}_N^\top \cdot \mathbf{H} \cdot \frac{1}{\sqrt{2}}\bar{\boldsymbol{\eta}}_N\right]\\ &\le \left(\sqrt{\mathbb{E}\left[\left(\frac{1}{\sqrt{2}}\bar{\boldsymbol{\eta}}_N^{\mathrm{bias}}\right)^\top \cdot \mathbf{H} \cdot \frac{1}{\sqrt{2}}\bar{\boldsymbol{\eta}}_N^{\mathrm{bias}}\right]} + \sqrt{\mathbb{E}\left[\left(\frac{1}{\sqrt{2}}\bar{\boldsymbol{\eta}}_N^{\mathrm{variance}}\right)^\top \cdot \mathbf{H} \cdot \frac{1}{\sqrt{2}}\bar{\boldsymbol{\eta}}_N^{\mathrm{variance}}\right]}\right)^2\\ &= \left(\sqrt{\frac{1}{2}\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\eta}}_N^{\mathrm{bias}} \otimes \bar{\boldsymbol{\eta}}_N^{\mathrm{bias}}]\rangle} + \sqrt{\frac{1}{2}\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\eta}}_N^{\mathrm{variance}} \otimes \bar{\boldsymbol{\eta}}_N^{\mathrm{variance}}]\rangle}\right)^2, \end{aligned}$$

where we use Cauchy–Schwarz inequality in the inequality such that for any vector $\mathbf{u}$ and $\mathbf{v}$, $\mathbb{E}\|\mathbf{u}+\mathbf{v}\|_{\mathbf{H}}^2 \le \left(\sqrt{\mathbb{E}\|\mathbf{u}\|_{\mathbf{H}}^2} + \sqrt{\mathbb{E}\|\mathbf{v}\|_{\mathbf{H}}^2}\right)^2$. ■

Lemma B.3 *Recall iterates (4.4) and (4.5). If the stepsize satisfies $\gamma \le 1/\lambda_1$, the bias error and variance error are upper bounded respectively as follows:*

$$\mathrm{bias} := \frac{1}{2}\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\eta}}_N^{\mathrm{bias}} \otimes \bar{\boldsymbol{\eta}}_N^{\mathrm{bias}}]\rangle \le \frac{1}{N^2}\sum_{t=0}^{N-1}\sum_{k=t}^{N-1}\big\langle (\mathbf{I}-\gamma\mathbf{H})^{k-t}\mathbf{H}, \mathbf{B}_t\big\rangle,$$

$$\mathrm{variance} := \frac{1}{2}\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\eta}}_N^{\mathrm{variance}} \otimes \bar{\boldsymbol{\eta}}_N^{\mathrm{variance}}] \le \frac{1}{N^2}\sum_{t=0}^{N-1}\sum_{k=t}^{N-1}\big\langle (\mathbf{I}-\gamma\mathbf{H})^{k-t}\mathbf{H}, \mathbf{C}_t\big\rangle.$$

Proof The proof will largely rely on the calculation in Jain et al. (2017b). Firstly, based on the definitions of $\boldsymbol{\eta}_t^{\mathrm{bias}}$ and $\boldsymbol{\eta}_t^{\mathrm{bias}}$ provided in (4.2) and (4.3), we have

$$\mathbb{E}[\boldsymbol{\eta}_t^{\mathrm{bias}}|\boldsymbol{\eta}_{t-1}^{\mathrm{bias}}] = \mathbb{E}[\mathbf{P}_t\boldsymbol{\eta}_{t-1}^{\mathrm{bias}}|\boldsymbol{\eta}_{t-1}^{\mathrm{bias}}] = (\mathbf{I}-\gamma\mathbf{H})\boldsymbol{\eta}_{t-1}^{\mathrm{bias}}. \tag{B.3}$$

$$\mathbb{E}[\boldsymbol{\eta}_t^{\mathrm{variance}}|\boldsymbol{\eta}_{t-1}^{\mathrm{variance}}] = \mathbb{E}[\mathbf{P}_t\boldsymbol{\eta}_{t-1}^{\mathrm{variance}} + \gamma\xi_t\mathbf{x}_t|\boldsymbol{\eta}_{t-1}^{\mathrm{variance}}] = (\mathbf{I}-\gamma\mathbf{H})\boldsymbol{\eta}_{t-1}^{\mathrm{variance}}. \tag{B.4}$$

Then regarding the quantity $\mathbb{E}[\bar{\boldsymbol{\eta}}_N^{\mathrm{bias}} \otimes \bar{\boldsymbol{\eta}}_N^{\mathrm{bias}}]$, we have

$$\begin{aligned}
&\mathbb{E}[\bar{\boldsymbol{\eta}}_N^{\mathrm{bias}} \otimes \bar{\boldsymbol{\eta}}_N^{\mathrm{bias}}]\\
&= \frac{1}{N^2}\cdot\bigg(\sum_{0\le k\le t\le N-1}\mathbb{E}[\boldsymbol{\eta}_t^{\mathrm{bias}}\otimes\boldsymbol{\eta}_k^{\mathrm{bias}}] + \sum_{0\le t<k\le N-1}\mathbb{E}[\boldsymbol{\eta}_t^{\mathrm{bias}}\otimes\boldsymbol{\eta}_k^{\mathrm{bias}}]\bigg)\\
&\preceq \frac{1}{N^2}\cdot\bigg(\sum_{0\le k\le t\le N-1}\mathbb{E}[\boldsymbol{\eta}_t^{\mathrm{bias}}\otimes\boldsymbol{\eta}_k^{\mathrm{bias}}] + \sum_{0\le t\le k\le N-1}\mathbb{E}[\boldsymbol{\eta}_t^{\mathrm{bias}}\otimes\boldsymbol{\eta}_k^{\mathrm{bias}}]\bigg)\\
&= \frac{1}{N^2}\cdot\bigg(\sum_{0\le k\le t\le N-1}(\mathbf{I}-\gamma\mathbf{H})^{t-k}\mathbb{E}[\boldsymbol{\eta}_k^{\mathrm{bias}}\otimes\boldsymbol{\eta}_k^{\mathrm{bias}}] + \sum_{0\le t\le k\le N-1}\mathbb{E}[\boldsymbol{\eta}_t^{\mathrm{bias}}\otimes\boldsymbol{\eta}_t^{\mathrm{bias}}](\mathbf{I}-\gamma\mathbf{H})^{k-t}\bigg)\\
&= \frac{1}{N^2}\cdot\sum_{t=0}^{N-1}\sum_{k=t}^{N-1}\Big((\mathbf{I}-\gamma\mathbf{H})^{k-t}\mathbb{E}[\boldsymbol{\eta}_t^{\mathrm{bias}}\otimes\boldsymbol{\eta}_t^{\mathrm{bias}}] + \mathbb{E}[\boldsymbol{\eta}_t^{\mathrm{bias}}\otimes\boldsymbol{\eta}_t^{\mathrm{bias}}](\mathbf{I}-\gamma\mathbf{H})^{k-t}\Big),
\end{aligned} \tag{B.5}$$

where we use (B.3) for $k-t$ (or $t-k$) times in the second equality. Therefore, plugging (B.5) into the inner product $\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\eta}}_N^{\mathrm{bias}} \otimes \bar{\boldsymbol{\eta}}_N^{\mathrm{bias}}]\rangle$ and noticing $\mathbf{H}$ is PSD, we have

$$\begin{aligned}
&\frac{1}{2}\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\eta}}_N^{\mathrm{bias}} \otimes \bar{\boldsymbol{\eta}}_N^{\mathrm{bias}}]\rangle\\
&\le \frac{1}{2N^2}\cdot\sum_{t=0}^{N-1}\sum_{k=t}^{N-1}\Big\langle \mathbf{H}, (\mathbf{I}-\gamma\mathbf{H})^{k-t}\mathbb{E}[\boldsymbol{\eta}_t^{\mathrm{bias}}\otimes\boldsymbol{\eta}_t^{\mathrm{bias}}] + \mathbb{E}[\boldsymbol{\eta}_t^{\mathrm{bias}}\otimes\boldsymbol{\eta}_t^{\mathrm{bias}}](\mathbf{I}-\gamma\mathbf{H})^{k-t}\Big\rangle\\
&= \frac{1}{N^2}\cdot\sum_{t=0}^{N-1}\sum_{k=t}^{N-1}\Big\langle (\mathbf{I}-\gamma\mathbf{H})^{k-t}\mathbf{H}, \mathbb{E}[\boldsymbol{\eta}_t^{\mathrm{bias}}\otimes\boldsymbol{\eta}_t^{\mathrm{bias}}]\Big\rangle
\end{aligned}$$

where the last equality holds since $\mathbf{H}$ and $(\mathbf{I}-\gamma\mathbf{H})^{k-t}$ commute.

By (B.4), we can similarly obtain the following for $\mathbb{E}[\bar{\boldsymbol{\eta}}_N^{\mathrm{variance}} \otimes \bar{\boldsymbol{\eta}}_N^{\mathrm{variance}}]$,

$$\begin{aligned}
&\mathbb{E}[\bar{\boldsymbol{\eta}}_N^{\mathrm{variance}} \otimes \bar{\boldsymbol{\eta}}_N^{\mathrm{variance}}]\\
&\preceq \frac{1}{N^2}\cdot\sum_{t=0}^{N-1}\sum_{k=t}^{N-1}\Big((\mathbf{I}-\gamma\mathbf{H})^{k-t}\mathbb{E}[\boldsymbol{\eta}_t^{\mathrm{variance}}\otimes\boldsymbol{\eta}_t^{\mathrm{variance}}] + \mathbb{E}[\boldsymbol{\eta}_t^{\mathrm{variance}}\otimes\boldsymbol{\eta}_t^{\mathrm{variance}}](\mathbf{I}-\gamma\mathbf{H})^{k-t}\Big),
\end{aligned}$$

which further leads to

$$\frac{1}{2}\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\eta}}_N^{\text{variance}} \otimes \bar{\boldsymbol{\eta}}_N^{\text{variance}}]\rangle \le \frac{1}{N^2} \cdot \sum_{t=0}^{N-1} \sum_{k=t}^{N-1} \Big\langle (\mathbf{I} - \gamma \mathbf{H})^{k-t}\mathbf{H}, \mathbb{E}[\boldsymbol{\eta}_t^{\text{variance}} \otimes \boldsymbol{\eta}_t^{\text{variance}}] \Big\rangle.$$

This completes the proof.

■

B.3 Bounding the Variance Error

We first introduce a weaker assumption (compared with Assumption 2.2) on the data distribution, which is sufficient to get our desired results on the variance error.

Assumption B.1 *There exists a constant $R > 0$ such that $\mathbb{E}[\mathbf{x}\mathbf{x}^\top \mathbf{x}\mathbf{x}^\top] \preceq R^2 \mathbf{H}$.*

We make this assumption to emphasize that our variance analysis does not rely on stronger assumptions than those in a number of prior works for iterate averaged SGD (Bach and Moulines, 2013; Jain et al., 2017b; Berthier et al., 2020). Moreover, note that this assumption is implied by Assumption 2.2 by setting $\mathbf{A} = \mathbf{I}$, which gives $R^2 = \alpha \operatorname{tr}(\mathbf{H})$.

Recall the variance error upper bound in Lemma B.3:

$$\text{variance} \le \frac{1}{N^2} \sum_{t=0}^{N-1} \sum_{k=t}^{N-1} \big\langle (\mathbf{I} - \gamma \mathbf{H})^{k-t}\mathbf{H}, \mathbf{C}_t \big\rangle.$$

We first have the following crude bound on $\mathbf{C}_t$.

Lemma B.4 *((Jain et al., 2017a) Lemma 5) Under Assumptions 2.1, 2.3 and B.1, if the stepsize satisfies $\gamma < 1/R^2$, it holds that*

$$0 = \mathbf{C}_0 \preceq \mathbf{C}_1 \preceq \cdots \preceq \mathbf{C}_\infty \preceq \frac{\gamma \sigma^2}{1 - \gamma R^2}\mathbf{I}.$$

Proof This lemma directly comes from Lemmas 3 and 5 in Jain et al. (2017a). For completeness, a proof is included as follows.

We first show that $\mathbf{C}_t$ is increasing:

$$\begin{aligned}
\mathbf{C}_t &= (\mathcal{I} - \gamma \mathcal{T}) \circ \mathbf{C}_{t-1} + \gamma^2 \boldsymbol{\Sigma} \\
&= \gamma^2 \sum_{k=0}^{t-1} (\mathcal{I} - \gamma \mathcal{T})^k \circ \boldsymbol{\Sigma} \qquad \text{(solving the recursion)} \\
&= \mathbf{C}_{t-1} + \gamma^2 (\mathcal{I} - \gamma \mathcal{T})^{t-1} \circ \boldsymbol{\Sigma} \\
&\succeq \mathbf{C}_{t-1}. \qquad \text{(since } \mathcal{I} - \gamma \mathcal{T} \text{ is a PSD mapping by Lemma 4.1)}
\end{aligned}$$

Next we show that $\mathbf{C}_\infty$ exists. Since $\mathbf{C}_t$ is PSD and increasing, it suffices to show that $\operatorname{tr}(\mathbf{C}_t)$ can be bounded uniformly. For any $t \ge 1$, we have

$$\mathbf{C}_t = \gamma^2 \sum_{k=0}^{t-1} (\mathcal{I} - \gamma \mathcal{T})^k \circ \boldsymbol{\Sigma} \preceq \gamma^2 \sum_{t=0}^{\infty} (\mathcal{I} - \gamma \mathcal{T})^t \circ \boldsymbol{\Sigma}. \tag{B.6}$$

Let $\mathbf{A}_t := (\mathcal{I} - \gamma\mathcal{T})^t \circ \mathbf{\Sigma}$, then $\mathbf{A}_t = (\mathcal{I} - \gamma\mathcal{T}) \circ \mathbf{A}_{t-1}$. By Assumption B.1 we have $\mathbb{E}[\mathbf{x}\mathbf{x}^\top\mathbf{x}\mathbf{x}^\top] \preceq R^2\mathbf{H}$. Then, by (B.1), we can get

$$\begin{aligned}\operatorname{tr}(\mathbf{A}_t) &= \operatorname{tr}(\mathbf{A}_{t-1}) - 2\gamma\operatorname{tr}(\mathbf{H}\mathbf{A}_{t-1}) + \gamma^2\operatorname{tr}\left(\mathbf{A}_{t-1}\mathbb{E}[\mathbf{x}\mathbf{x}^\top\mathbf{x}\mathbf{x}^\top]\right)\\ &\le \operatorname{tr}(\mathbf{A}_{t-1}) - (2\gamma - \gamma^2R^2)\operatorname{tr}(\mathbf{H}\mathbf{A}_{t-1})\\ &\le \operatorname{tr}\left((\mathbf{I} - \gamma\mathbf{H})\mathbf{A}_{t-1}\right)\\ &\le (1 - \gamma\lambda_d)\operatorname{tr}(\mathbf{A}_{t-1}), \end{aligned} \tag{B.7}$$

where we use the assumption $\gamma \le 1/R^2$ in the second inequality. Combining (B.6) and (B.7), we have for any $t \ge 1$ that

$$\operatorname{tr}(\mathbf{C}_t) \le \gamma^2\sum_{t=0}^{\infty}\operatorname{tr}\left((\mathcal{I} - \gamma\mathcal{T})^t \circ \mathbf{\Sigma}\right) = \gamma^2\sum_{t=0}^{\infty}\operatorname{tr}(\mathbf{A}_t) \le \frac{\gamma\operatorname{tr}(\mathbf{\Sigma})}{\lambda_d} < \infty.$$

Therefore, $\operatorname{tr}(\mathbf{C}_t)$ is uniformly upper bounded, hence $\mathbf{C}_\infty$ exists.

Finally we upper bound $\mathbf{C}_\infty$. Taking limits in (4.4), we have

$$\mathbf{C}_\infty = (\mathcal{I} - \gamma\mathcal{T}) \circ \mathbf{C}_\infty + \gamma^2\mathbf{\Sigma},$$

which immediately implies

$$\mathbf{C}_\infty = \gamma\mathcal{T}^{-1} \circ \mathbf{\Sigma}.$$

Recalling $\widetilde{\mathcal{T}} = \mathcal{T} + \gamma\mathcal{M} - \gamma\widetilde{\mathcal{M}}$ and the definitions and properties of the operators, we have

$$\begin{aligned}\widetilde{\mathcal{T}} \circ \mathbf{C}_\infty &= \mathcal{T} \circ \mathbf{C}_\infty + \gamma\mathcal{M} \circ \mathbf{C}_\infty - \gamma\widetilde{\mathcal{M}} \circ \mathbf{C}_\infty\\ &= \gamma\mathbf{\Sigma} + \gamma\mathcal{M} \circ \mathbf{C}_\infty - \gamma\widetilde{\mathcal{M}} \circ \mathbf{C}_\infty \qquad (\text{since } \mathbf{C}_\infty = \gamma\mathcal{T}^{-1} \circ \mathbf{\Sigma})\\ &\preceq \gamma\mathbf{\Sigma} + \gamma\mathcal{M} \circ \mathbf{C}_\infty \qquad (\text{since } \widetilde{\mathcal{M}} \text{ is a PSD mapping by Lemma 4.1})\\ &\preceq \gamma\sigma^2\mathbf{H} + \gamma\mathcal{M} \circ \mathbf{C}_\infty. \qquad (\text{since } \mathbf{\Sigma} \preceq \sigma^2\mathbf{H} \text{ by Assumption 2.3})\end{aligned}$$

Recall that $\widetilde{\mathcal{T}}^{-1}$ exists and is a PSD mapping by Lemma 4.1, we then have

$$\begin{aligned}\mathbf{C}_\infty &\preceq \gamma\sigma^2 \cdot \widetilde{\mathcal{T}}^{-1} \circ \mathbf{H} + \gamma\widetilde{\mathcal{T}}^{-1} \circ \mathcal{M} \circ \mathbf{C}_\infty\\ &\preceq \gamma\sigma^2 \cdot \sum_{t=0}^{\infty}(\gamma\widetilde{\mathcal{T}}^{-1} \circ \mathcal{M})^t \circ \widetilde{\mathcal{T}}^{-1} \circ \mathbf{H}. \qquad (\text{solving the recursion})\end{aligned} \tag{B.8}$$

In addition, we have

$$\begin{aligned}\widetilde{\mathcal{T}}^{-1} \circ \mathbf{H} &= \gamma\sum_{t=0}^{\infty}(\mathcal{I} - \gamma\widetilde{\mathcal{T}})^t \circ \mathbf{H}\\ &= \gamma\sum_{t=0}^{\infty}(\mathbf{I} - \gamma\mathbf{H})^t\mathbf{H}(\mathbf{I} - \gamma\mathbf{H})^t \qquad (\text{by the property of } \mathcal{I} - \widetilde{\mathcal{T}} \text{ in (4.1)})\\ &\preceq \gamma\sum_{t=0}^{\infty}(\mathbf{I} - \gamma\mathbf{H})^t\mathbf{H}\end{aligned}$$

$$= \mathbf{I}. \tag{B.9}$$

Substituting (B.9) into (B.8), we obtain

$$\begin{aligned}
\mathbf{C}_\infty &\preceq \gamma\sigma^2 \cdot \sum_{t=0}^{\infty} (\gamma\widetilde{\mathcal{T}}^{-1} \circ \mathcal{M})^t \circ \mathbf{I} \\
&= \gamma\sigma^2 \cdot \sum_{t=0}^{\infty} (\gamma\widetilde{\mathcal{T}}^{-1} \circ \mathcal{M})^{t-1} \circ \gamma\widetilde{\mathcal{T}}^{-1} \circ \mathcal{M} \circ \mathbf{I} \\
&\preceq \gamma\sigma^2 \cdot \sum_{t=0}^{\infty} (\gamma\widetilde{\mathcal{T}}^{-1} \circ \mathcal{M})^{t-1} \circ \gamma R^2 \mathbf{I} \\
&\preceq \gamma\sigma^2 \cdot \sum_{t=0}^{\infty} (\gamma R^2)^t \mathbf{I} \\
&= \frac{\gamma\sigma^2}{1-\gamma R^2} \mathbf{I},
\end{aligned}$$

where the second inequality is due to $\mathcal{M} \circ \mathbf{I} \preceq R^2\mathbf{H}$ by Assumption B.1 and $\widetilde{\mathcal{T}}^{-1} \circ \mathbf{H} \preceq \mathbf{I}$ in (B.9), and the third inequality is by recursion. This completes the proof. ■

The following lemma refines the bound on $\mathbf{C}_t$ by its update rule and its crude bound shown in previous lemma.

Lemma B.5 *Under Assumptions 2.1, 2.3 and B.1, if the stepsize satisfies $\gamma < 1/R^2$, it holds that*

$$\mathbf{C}_t \preceq \frac{\gamma\sigma^2}{1-\gamma R^2} \cdot \big(\mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^t\big).$$

Proof By (4.5) and the definitions of $\mathcal{T}$ and $\widetilde{\mathcal{T}}$, we have

$$\begin{aligned}
\mathbf{C}_t &= (\mathcal{I} - \gamma\mathcal{T}) \circ \mathbf{C}_{t-1} + \gamma^2\boldsymbol{\Sigma} \\
&= (\mathcal{I} - \gamma\widetilde{\mathcal{T}}) \circ \mathbf{C}_{t-1} + \gamma^2(\mathcal{M} - \widetilde{\mathcal{M}}) \circ \mathbf{C}_{t-1} + \gamma^2\boldsymbol{\Sigma} \\
&\preceq (\mathcal{I} - \gamma\widetilde{\mathcal{T}}) \circ \mathbf{C}_{t-1} + \gamma^2\mathcal{M} \circ \mathbf{C}_{t-1} + \gamma^2\boldsymbol{\Sigma},
\end{aligned} \tag{B.10}$$

where the last inequality is due to the fact that $\widetilde{\mathcal{M}}$ is a PSD mapping. Then by Lemma B.4, we have for all $t \ge 0$,

$$\mathcal{M} \circ \mathbf{C}_t \preceq \mathcal{M} \circ \mathbf{C}_\infty \preceq \mathcal{M} \circ \frac{\gamma\sigma^2}{1-\gamma R^2}\mathbf{I} = \frac{\gamma\sigma^2}{1-\gamma R^2} \cdot \mathbb{E}[\|\mathbf{x}\|_2^2\,\mathbf{x}\mathbf{x}^\top] \preceq \frac{\gamma R^2\sigma^2}{1-\gamma R^2} \cdot \mathbf{H}. \tag{B.11}$$

Substituting (B.11) and $\boldsymbol{\Sigma} \preceq \big\|\mathbf{H}^{-1/2}\boldsymbol{\Sigma}\mathbf{H}^{-1/2}\big\|_2 \cdot \mathbf{H}$ into (B.10), we obtain

$$\begin{aligned}
\mathbf{C}_t &\preceq (\mathcal{I} - \gamma\widetilde{\mathcal{T}}) \circ \mathbf{C}_{t-1} + \gamma^2 \cdot \frac{\gamma R^2\sigma^2}{1-\gamma R^2} \cdot \mathbf{H} + \gamma^2 \cdot \|\mathbf{H}^{-1/2}\boldsymbol{\Sigma}\mathbf{H}^{-1/2}\|_2 \cdot \mathbf{H} \\
&= (\mathcal{I} - \gamma\widetilde{\mathcal{T}}) \circ \mathbf{C}_{t-1} + \gamma^2 \cdot \frac{\gamma R^2\sigma^2}{1-\gamma R^2} \cdot \mathbf{H} + \gamma^2\sigma^2 \cdot \mathbf{H}
\end{aligned}$$

$$
\begin{aligned}
&= (\mathcal{I} - \gamma\widetilde{\mathcal{T}}) \circ \mathbf{C}_{t-1} + \frac{\gamma^2\sigma^2}{1-\gamma R^2} \cdot \mathbf{H} \\
&\preceq \frac{\gamma^2\sigma^2}{1-\gamma R^2} \cdot \sum_{k=0}^{t-1} (\mathcal{I} - \gamma\widetilde{\mathcal{T}})^k \circ \mathbf{H}. \qquad \text{(solving the recursion)} \\
&= \frac{\gamma^2\sigma^2}{1-\gamma R^2} \cdot \sum_{k=0}^{t-1} (\mathbf{I} - \gamma\mathbf{H})^k \mathbf{H} (\mathbf{I} - \gamma\mathbf{H})^k \qquad \text{(by the property of } \mathcal{I} - \gamma\widetilde{\mathcal{T}} \text{ in (4.1))} \\
&\preceq \frac{\gamma^2\sigma^2}{1-\gamma R^2} \cdot \sum_{k=0}^{t-1} (\mathbf{I} - \gamma\mathbf{H})^k \mathbf{H} \\
&= \frac{\gamma\sigma^2}{1-\gamma R^2} \cdot \left(\mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^t\right),
\end{aligned}
$$

where in the last inequality we use $\gamma \le 1/R^2 \le 1/\operatorname{tr}(\mathbf{H}) \le 1/\lambda_1$. This completes the proof. ■

We are ready to provide the variance error upper bound.

Lemma B.6 *Under Assumptions 2.1, 2.3 and B.1, if the stepsize satisfies $\gamma < 1/R^2$, then it holds that*

$$
\text{variance} \le \frac{\sigma^2}{1-\gamma R^2} \left(\frac{k^*}{N} + \gamma^2 N \cdot \sum_{i>k^*} \lambda_i^2 \right),
$$

where $k^* = \max\{k : \lambda_k \ge \frac{1}{N\gamma}\}$.

Proof By Lemma B.2, we can bound the variance error as follows

$$
\begin{aligned}
\text{variance} &\le \frac{1}{N^2} \sum_{t=0}^{N-1} \sum_{k=t}^{N-1} \left\langle (\mathbf{I} - \gamma\mathbf{H})^{k-t}\mathbf{H}, \mathbf{C}_t \right\rangle \\
&= \frac{1}{\gamma N^2} \sum_{t=0}^{N-1} \left\langle \mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^{N-t}, \mathbf{C}_t \right\rangle \\
&\le \frac{\sigma^2}{N^2(1-\gamma R^2)} \sum_{t=0}^{N-1} \left\langle \mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^{N-t}, \left(\mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^t\right) \right\rangle \\
&= \frac{\sigma^2}{N^2(1-\gamma R^2)} \sum_i \sum_{t=0}^{N-1} \left(1 - (1-\gamma\lambda_i)^{N-t}\right)\left(1 - (1-\gamma\lambda_i)^t\right) \\
&\le \frac{\sigma^2}{N^2(1-\gamma R^2)} \sum_i \sum_{t=0}^{N-1} \left(1 - (1-\gamma\lambda_i)^{N}\right)\left(1 - (1-\gamma\lambda_i)^N\right) \\
&= \frac{\sigma^2}{N(1-\gamma R^2)} \sum_i \left(1 - (1-\gamma\lambda_i)^N\right)^2,
\end{aligned}
$$

where the second inequality is due to Lemma B.5, $\{\lambda_i\}_{i\ge 1}$ are the eigenvalues of $\mathbf{H}$ and are sorted in decreasing order. Since $\gamma \le 1/\lambda_1$, we have for all $i \ge 1$ that

$$
1 - (1-\gamma\lambda_i)^N \le \min\left\{1, \gamma N \lambda_i\right\}. \tag{B.12}
$$

Set $k^* = \max\{k : \lambda_k \geq \frac{1}{\gamma N}\}$, then

$$\begin{aligned}\text{variance} &\leq \frac{\sigma^2}{N(1-\gamma R^2)} \sum_i \min\left\{1, \gamma^2 N^2 \lambda_i^2\right\} \\ &\leq \frac{\sigma^2}{N(1-\gamma R^2)} \left(k^* + N^2\gamma^2 \cdot \sum_{i>k^*} \lambda_i^2\right) \\ &= \frac{\sigma^2}{1-\gamma R^2} \left(\frac{k^*}{N} + \gamma^2 N \cdot \sum_{i>k^*} \lambda_i^2\right).\end{aligned}$$

■

B.4 Bounding the Bias Error

In this part we will focus on bounding the bias error. Recall the bias error bound in Lemma B.3:

$$\begin{aligned}\text{bias} &\leq \frac{1}{N^2} \sum_{t=0}^{N-1} \sum_{k=t}^{N-1} \left\langle (\mathbf{I} - \gamma\mathbf{H})^{k-t}\mathbf{H}, \mathbf{B}_t \right\rangle \\ &= \frac{1}{\gamma N^2} \sum_{t=0}^{N-1} \left\langle \mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^{N-t}, \mathbf{B}_t \right\rangle \\ &\leq \frac{1}{\gamma N^2} \big\langle \mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^{N}, \sum_{t=0}^{N-1} \mathbf{B}_t \big\rangle. \end{aligned} \tag{B.13}$$

Let $\mathbf{S}_n = \sum_{t=0}^{n-1} \mathbf{B}_t$, then we only need to bound $\mathbf{S}_N$.

Lemma B.7 *Let* $\mathbf{S}_t = \sum_{k=0}^{t-1} \mathbf{B}_k$*, if* $\gamma < 1/(\alpha \operatorname{tr}(\mathbf{A}))$*, we have*

$$\mathbf{S}_t = (\mathcal{I} - \gamma\mathcal{T}) \circ \mathbf{S}_{t-1} + \mathbf{B}_0.$$

Moreover, it holds that

$$\mathbf{B}_0 = \mathbf{S}_0 \preceq \mathbf{S}_1 \preceq \cdots \preceq \mathbf{S}_\infty.$$

Proof By (4.4), we have

$$\mathbf{B}_t = (\mathcal{I} - \gamma\mathcal{T}) \circ \mathbf{B}_{t-1} = (\mathcal{I} - \gamma\mathcal{T})^t \circ \mathbf{B}_0, \tag{B.14}$$

where we used recursion. Then we have

$$\mathbf{S}_t = \sum_{k=0}^{t-1} (\mathcal{I} - \gamma\mathcal{T})^k \circ \mathbf{B}_0 = (\mathcal{I} - \gamma\mathcal{T}) \circ \left(\sum_{k=0}^{t-1} (\mathcal{I} - \gamma\mathcal{T})^k \circ \mathbf{B}_0\right) + \mathbf{B}_0 = (\mathcal{I} - \gamma\mathcal{T}) \circ \mathbf{S}_{t-1} + \mathbf{B}_0.$$

Moreover, since $\mathbf{B}_t$ is PSD for all $t \geq 0$, it is clear that $\mathbf{S}_t = \mathbf{S}_{t-1} + \mathbf{B}_t \succeq \mathbf{S}_{t-1}$. Besides, by Lemma 4.1, we know that

$$\mathbf{S}_\infty := \sum_{k=0}^{\infty} (\mathcal{I} - \gamma \mathcal{T})^t \circ \mathbf{B}_0 = \gamma^{-1} \mathcal{T}^{-1} \circ \mathbf{B}_0$$

exists. Thus it can be readily shown that

$$\mathbf{B}_0 = \mathbf{S}_1 \preceq \cdots \preceq \mathbf{S}_t \preceq \mathbf{S}_{t+1} \preceq \cdots \preceq \mathbf{S}_\infty,$$

which completes the proof. ■

Lemma B.8 *Under Assumptions 2.2, for any symmetric matrix* $\mathbf{A}$*, if* $\gamma < 1/(\alpha \operatorname{tr}(\mathbf{H}))$*, it holds that*

$$\mathcal{M} \circ \mathcal{T}^{-1} \circ \mathbf{A} \preceq \frac{\alpha \operatorname{tr}(\mathbf{A})}{1 - \gamma \alpha \operatorname{tr}(\mathbf{H})} \cdot \mathbf{H}.$$

Proof We first tackle $\mathcal{T}^{-1} \circ \mathbf{A}$. In particular, by Lemma 4.1 we have the operator $\mathcal{T}^{-1}$ exists and thus $\mathcal{T}^{-1} \circ \mathbf{A}$ also exists, which can be obtained by solving for the PSD matrix $\mathbf{D}$ satisfying the following equation,

$$\mathcal{T} \circ \mathbf{D} = \mathbf{A}.$$

Using the definition of $\widetilde{\mathcal{T}}$, we have:

$$\widetilde{\mathcal{T}} \circ \mathbf{D} = \gamma \mathcal{M} \circ \mathbf{D} + \mathbf{A} - \gamma \mathbf{H}\mathbf{D}\mathbf{H}, \tag{B.15}$$

where $\mathcal{M} \circ \mathbf{D} = \mathbb{E}[\mathbf{x}\mathbf{x}^\top \mathbf{D} \mathbf{x}\mathbf{x}^\top]$. Further by Lemma 4.1 we know that $\widetilde{\mathcal{T}}^{-1}$ and $\mathcal{M}$ are both PSD mapping. This implies that for any PSD matrices $\mathbf{U}$ and $\mathbf{U}'$ satisfying $\mathbf{0} \preceq \mathbf{U} \preceq \mathbf{U}'$, it holds that

$$\mathbf{0} \preceq \mathcal{M} \circ \mathbf{U} \preceq \mathcal{M} \circ \mathbf{U}', \qquad \mathbf{0} \preceq \widetilde{\mathcal{T}}^{-1} \circ \mathbf{U} \preceq \widetilde{\mathcal{T}}^{-1} \circ \mathbf{U}'.$$

Combining the above two results we also have

$$\mathbf{0} \preceq \mathcal{M} \circ \widetilde{\mathcal{T}}^{-1} \circ \mathbf{U} \preceq \mathcal{M} \circ \widetilde{\mathcal{T}}^{-1} \circ \mathbf{U}'. \tag{B.16}$$

Therefore, applying the operator $\mathcal{T}^{-1}$ to both sides of (B.15) yields

$$\begin{aligned} \mathbf{D} &= \gamma \widetilde{\mathcal{T}}^{-1} \circ \mathcal{M} \circ \mathbf{D} + \widetilde{\mathcal{T}}^{-1} \circ \mathbf{A} - \gamma \widetilde{\mathcal{T}}^{-1} \circ (\mathbf{H}\mathbf{D}\mathbf{H}) \\ &\preceq \gamma \widetilde{\mathcal{T}}^{-1} \circ \mathcal{M} \circ \mathbf{D} + \widetilde{\mathcal{T}}^{-1} \circ \mathbf{A}. \end{aligned} \tag{B.17}$$

Then we can apply the operator $\mathcal{M}$ to both sides of (B.17), by the monotonicity property in (B.16), we have

$$\mathcal{M} \circ \mathbf{D} \preceq \gamma \mathcal{M} \circ \widetilde{\mathcal{T}}^{-1} \circ \mathcal{M} \circ \mathbf{D} + \mathcal{M} \circ \widetilde{\mathcal{T}}^{-1} \circ \mathbf{A}$$

$$\preceq \sum_{t=0}^{\infty}(\gamma\mathcal{M}\circ\widetilde{\mathcal{T}}^{-1})^t\circ(\mathcal{M}\circ\widetilde{\mathcal{T}}^{-1}\circ\mathbf{A}). \tag{B.18}$$

By Assumption 2.2 we have

$$\mathcal{M}\circ\widetilde{\mathcal{T}}^{-1}\circ\mathbf{A}\preceq\alpha\,\mathrm{tr}(\mathbf{H}\widetilde{\mathcal{T}}^{-1}\circ\mathbf{A})\mathbf{H}. \tag{B.19}$$

Additionally, based on the definition of $\widetilde{\mathcal{T}}$, we have

$$\widetilde{\mathcal{T}}^{-1}\circ\mathbf{A}=\gamma\sum_{t=0}^{\infty}(\mathcal{I}-\gamma\widetilde{\mathcal{T}})^t\circ\mathbf{A}=\gamma\sum_{t=0}^{\infty}(\mathbf{I}-\gamma\mathbf{H})^t\mathbf{A}(\mathbf{I}-\gamma\mathbf{H})^t.$$

Therefore, it follows that

$$\begin{aligned}\mathrm{tr}(\mathbf{H}\widetilde{\mathcal{T}}^{-1}\circ\mathbf{A})&=\gamma\,\mathrm{tr}\bigg(\sum_{t=0}^{\infty}\mathbf{H}(\mathbf{I}-\gamma\mathbf{H})^t\mathbf{A}(\mathbf{I}-\gamma\mathbf{H})^t\bigg)\\&=\gamma\,\mathrm{tr}\bigg(\sum_{t=0}^{\infty}\mathbf{H}(\mathbf{I}-\gamma\mathbf{H})^{2t}\mathbf{A}\bigg)\\&=\mathrm{tr}\big(\mathbf{H}(2\mathbf{H}-\gamma\mathbf{H}^2)^{-1}\mathbf{A}\big)\\&\le\mathrm{tr}(\mathbf{A}),\end{aligned} \tag{B.20}$$

where the last inequality is because we have $\gamma\le 1/\lambda_1$ and thus $\mathbf{H}(2\mathbf{H}-\gamma\mathbf{H}^2)^{-1}\preceq\mathbf{I}$. Substituting (B.20) into (B.19) yields

$$\mathcal{M}\circ\widetilde{\mathcal{T}}^{-1}\circ\mathbf{A}\preceq\alpha\,\mathrm{tr}(\mathbf{A})\mathbf{H}.$$

Note that we have $\widetilde{\mathcal{T}}^{-1}\mathbf{H}\preceq\mathbf{I}$ and $\mathcal{M}\circ\mathbf{I}\preceq\alpha\,\mathrm{tr}(\mathbf{H})\mathbf{H}$, plugging the above inequality into (B.18) gives

$$\mathcal{M}\circ\mathcal{T}^{-1}\circ\mathbf{A}=\mathcal{M}\circ\mathbf{D}\preceq\alpha\,\mathrm{tr}(\mathbf{A})\sum_{t=0}^{\infty}(\gamma\alpha\,\mathrm{tr}(\mathbf{H}))^t\mathbf{H}\preceq\frac{\alpha\,\mathrm{tr}(\mathbf{A})}{1-\gamma\alpha\,\mathrm{tr}(\mathbf{H})}\cdot\mathbf{H}.$$

This completes the proof. ■

Lemma B.9 *Under Assumptions 2.1, and 2.2, if the stepsize satisfies $\gamma<1/(\alpha\,\mathrm{tr}(\mathbf{H}))$, then*

$$\mathcal{M}\circ\mathbf{S}_t\preceq\frac{\alpha\cdot\mathrm{tr}\big(\big[\mathcal{I}-(\mathcal{I}-\gamma\widetilde{\mathcal{T}})^t\big]\circ\mathbf{B}_0\big)}{\gamma(1-\gamma\alpha\,\mathrm{tr}(\mathbf{H}))}\cdot\mathbf{H}.$$

Proof Note that $\mathbf{S}_t$ takes the following form

$$\mathbf{S}_t:=\sum_{k=0}^{t-1}(\mathcal{I}-\gamma\mathcal{T})^k\circ\mathbf{B}_0=\gamma^{-1}\mathcal{T}^{-1}\circ\big[\mathcal{I}-(\mathcal{I}-\gamma\mathcal{T})^t\big]\mathbf{B}_0.$$

Note that by Lemma 4.1, we have $\mathcal{I} - \gamma\widetilde{\mathcal{T}} \le \mathcal{I} - \gamma\mathcal{T}$ so that $\mathcal{I} - (\mathcal{I} - \gamma\mathcal{T})^t \preceq \mathcal{I} - (\mathcal{I} - \gamma\widetilde{\mathcal{T}})^t$. Therefore, further note that $\mathcal{T}^{-1}$ is a PSD mapping, we have the following bound on $\mathbf{S}_t$,

$$\mathbf{S}_t \preceq \gamma^{-1}\mathcal{T}^{-1} \circ \left[\mathcal{I} - (\mathcal{I} - \gamma\widetilde{\mathcal{T}})^t\right] \circ \mathbf{B}_0.$$

Then note that $\left[\mathcal{I} - (\mathcal{I} - \gamma\widetilde{\mathcal{T}})^t\right] \circ \mathbf{B}_0$ is a PSD matrix, applying Lemma B.8, we get

$$\mathcal{M} \circ \mathbf{S}_t \preceq \gamma^{-1}\mathcal{M} \circ \mathcal{T}^{-1} \circ \left[\mathcal{I} - (\mathcal{I} - \gamma\widetilde{\mathcal{T}})^t\right] \circ \mathbf{B}_0 \preceq \frac{\alpha \cdot \operatorname{tr}\left(\left[\mathcal{I} - (\mathcal{I} - \gamma\widetilde{\mathcal{T}})^t\right] \circ \mathbf{B}_0\right)}{\gamma(1 - \gamma\alpha\operatorname{tr}(\mathbf{H}))} \cdot \mathbf{H}.$$

This completes the proof. ■

The following lemma shows that using this crude bound on $\mathcal{M} \circ \mathbf{S}_t$ we are able to get a tighter upper bound on $\mathbf{S}_t$.

Lemma B.10 *Under Assumptions 2.1 and 2.2, let* $\mathbf{B}_{a,b} = \mathbf{B}_a - (\mathbf{I} - \gamma\mathbf{H})^{b-a}\mathbf{B}_a(\mathbf{I} - \gamma\mathbf{H})^{b-a}$, *if the stepsize satisfies* $\gamma < 1/(\alpha\operatorname{tr}(\mathbf{H}))$, *then for any* $t \le N$, *it holds that*

$$\mathbf{S}_t \preceq \sum_{k=0}^{t-1}(\mathbf{I} - \gamma\mathbf{H})^k\left(\frac{\gamma\alpha\operatorname{tr}(\mathbf{B}_{0,N})}{1 - \gamma\alpha\operatorname{tr}(\mathbf{H})} \cdot \mathbf{H} + \mathbf{B}_0\right)(\mathbf{I} - \gamma\mathbf{H})^k.$$

Proof Recall the recursive form of $\mathbf{S}_t$ given in Lemma B.7, we have

$$\mathbf{S}_t = (\mathcal{I} - \gamma\mathcal{T}) \circ \mathbf{S}_{t-1} + \mathbf{B}_0.$$

Note that this is similar to the recursive form of $\mathbf{C}_t$ provided in (4.5) but replacing $\gamma^2\mathbf{\Sigma}$ with $\mathbf{B}_0$. Then we can use the similar proof of Lemma B.5 to get the upper bound of $\mathbf{S}_t$. In particular, note that we will run SGD with N steps, then $\mathbf{S}_N$ can be used as a uniform upper bound on $\mathbf{S}_1, \ldots, \mathbf{S}_N$, we can upper bound $\mathbf{S}_t$ by

$$\begin{aligned}
\mathbf{S}_t &\preceq (\mathcal{I} - \gamma\widetilde{\mathcal{T}}) \circ \mathbf{S}_{t-1} + \gamma^2\mathcal{M} \circ \mathbf{S}_N + \mathbf{B}_0 \\
&\preceq (\mathcal{I} - \gamma\widetilde{\mathcal{T}}) \circ \mathbf{S}_{t-1} + \frac{\gamma\alpha \cdot \operatorname{tr}\left(\left[\mathcal{I} - (\mathcal{I} - \gamma\widetilde{\mathcal{T}})^N\right] \circ \mathbf{B}_0\right)}{1 - \gamma\alpha\operatorname{tr}(\mathbf{H})} \cdot \mathbf{H} + \mathbf{B}_0 \\
&= \sum_{k=0}^{t-1}(\mathcal{I} - \gamma\widetilde{\mathcal{T}})^k \circ \left(\frac{\gamma\alpha \cdot \operatorname{tr}\left(\left[\mathcal{I} - (\mathcal{I} - \gamma\widetilde{\mathcal{T}})^N\right] \circ \mathbf{B}_0\right)}{1 - \gamma\alpha\operatorname{tr}(\mathbf{H})} \cdot \mathbf{H} + \mathbf{B}_0\right) \\
&= \sum_{k=0}^{t-1}(\mathbf{I} - \gamma\mathbf{H})^k\left(\frac{\gamma\alpha\operatorname{tr}\left(\mathbf{B}_0 - (\mathbf{I} - \gamma\mathbf{H})^N\mathbf{B}_0(\mathbf{I} - \gamma\mathbf{H})^N\right)}{1 - \gamma\alpha\operatorname{tr}(\mathbf{H})} \cdot \mathbf{H} + \mathbf{B}_0\right)(\mathbf{I} - \gamma\mathbf{H})^k.
\end{aligned}$$

where we use Lemma B.9 in the second inequality, the first equality is by recursion, and the last equality is by the definition of $\widetilde{\mathcal{T}}$. ■

We now put these lemmas together and provide our upper bound on the bias error:

Lemma B.11 *Under Assumptions 2.1 and 2.2, if the stepsize satisfies* $\gamma < 1/(\alpha \operatorname{tr}(\mathbf{H}))$, *it holds that*

$$\begin{aligned}\mathrm{bias} \le &\frac{1}{\gamma^2 N^2}\cdot \|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k^*}} + \|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}} \\ &+ \frac{2\alpha\big(\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{I}_{0:k^*}} + N\gamma\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}}\big)}{\gamma N(1-\gamma\alpha\operatorname{tr}(\mathbf{H}))}\cdot\bigg(\frac{k^*}{N} + N\gamma^2\sum_{i>k^*}\lambda_i^2\bigg),\end{aligned}$$

where $k^* = \max\{k : \lambda_k \ge \gamma^{-1}/N\}$.

Proof We can plug the upper bound of $\mathbf{S}_t$ derived in Lemma B.10 into (B.13) and get

$$\begin{aligned}\mathrm{bias} &\le \frac{1}{\gamma N^2}\sum_{k=0}^{N-1}\bigg\langle \mathbf{I} - (\mathbf{I}-\gamma\mathbf{H})^N, (\mathbf{I}-\gamma\mathbf{H})^k\bigg(\frac{\gamma\alpha\operatorname{tr}(\mathbf{B}_{0,N})}{1-\gamma\alpha\operatorname{tr}(\mathbf{H})}\cdot\mathbf{H} + \mathbf{B}_0\bigg)(\mathbf{I}-\gamma\mathbf{H})^k\bigg\rangle \\ &= \frac{1}{\gamma N^2}\sum_{k=0}^{N-1}\bigg\langle (\mathbf{I}-\gamma\mathbf{H})^{2k} - (\mathbf{I}-\gamma\mathbf{H})^{N+2k}, \frac{\gamma\alpha\operatorname{tr}(\mathbf{B}_{0,N})}{1-\gamma\alpha\operatorname{tr}(\mathbf{H})}\cdot\mathbf{H} + \mathbf{B}_0\bigg\rangle.\end{aligned}$$

Note that

$$\begin{aligned}(\mathbf{I}-\gamma\mathbf{H})^{2k} - (\mathbf{I}-\gamma\mathbf{H})^{N+2k} &= (\mathbf{I}-\gamma\mathbf{H})^k\big((\mathbf{I}-\gamma\mathbf{H})^k - (\mathbf{I}-\gamma\mathbf{H})^{N+k}\big) \\ &\preceq (\mathbf{I}-\gamma\mathbf{H})^k - (\mathbf{I}-\gamma\mathbf{H})^{N+k}.\end{aligned}$$

We obtain

$$\mathrm{bias} \le \frac{1}{\gamma N^2}\sum_{k=0}^{N-1}\bigg\langle (\mathbf{I}-\gamma\mathbf{H})^k - (\mathbf{I}-\gamma\mathbf{H})^{N+k}, \frac{\gamma\alpha\operatorname{tr}(\mathbf{B}_{0,N})}{1-\gamma\alpha\operatorname{tr}(\mathbf{H})}\cdot\mathbf{H} + \mathbf{B}_0\bigg\rangle,$$

Therefore, it suffices to upper bound the following two terms:

$$\begin{aligned}I_1 &= \frac{\alpha\operatorname{tr}(\mathbf{B}_{0,N})}{N^2(1-\gamma\alpha\operatorname{tr}(\mathbf{H}))}\sum_{k=0}^{N-1}\big\langle (\mathbf{I}-\gamma\mathbf{H})^k - (\mathbf{I}-\gamma\mathbf{H})^{N+k}, \mathbf{H}\big\rangle \\ I_2 &= \frac{1}{\gamma N^2}\sum_{k=0}^{N-1}\big\langle (\mathbf{I}-\gamma\mathbf{H})^k - (\mathbf{I}-\gamma\mathbf{H})^{N+k}, \mathbf{B}_0\big\rangle.\end{aligned}$$

Regarding I_1, since $\mathbf{H}$ and $\mathbf{I}-\gamma\mathbf{H}$ can be diagonalized simultaneously, we have

$$\begin{aligned}I_1 &= \frac{\alpha\operatorname{tr}(\mathbf{B}_{0,N})}{N^2(1-\gamma\alpha\operatorname{tr}(\mathbf{H}))}\sum_{k=0}^{N-1}\sum_i\big[(1-\gamma\lambda_i)^k - (1-\gamma\lambda_i)^{N+k}\big]\lambda_i \\ &= \frac{\alpha\operatorname{tr}(\mathbf{B}_{0,N})}{\gamma N^2(1-\gamma\alpha\operatorname{tr}(\mathbf{H}))}\sum_i\big[1-(1-\gamma\lambda_i)^N\big]^2 \\ &\le \frac{\alpha\operatorname{tr}(\mathbf{B}_{0,N})}{\gamma N^2(1-\gamma\alpha\operatorname{tr}(\mathbf{H}))}\sum_i\min\big\{1, \gamma^2N^2\lambda_i^2\big\}\end{aligned}$$

$$\leq \frac{\alpha\,\mathrm{tr}(\mathbf{B}_{0,N})}{\gamma(1-\gamma\alpha\,\mathrm{tr}(\mathbf{H}))}\cdot\bigg(\frac{k^*}{N^2}+\gamma^2\sum_{i>k^*}\lambda_i^2\bigg), \tag{B.21}$$

where k^* is the index of the smallest eigenvalue of $\mathbf{H}$ satisfying $\lambda_{k^*}\geq\gamma^{-1}/N$. Moreover, recall that $\widetilde{\mathbf{B}}=\mathbf{B}_0-(\mathbf{I}-\gamma\mathbf{H})^N\mathbf{B}_0(\mathbf{I}-\gamma\mathbf{H})^N)$ and $\mathbf{B}_0=(\mathbf{w}_0-\mathbf{w}^*)\otimes(\mathbf{w}_0-\mathbf{w}^*)$, we have

$$\mathrm{tr}(\mathbf{B}_{0,N})=\mathrm{tr}\left(\mathbf{B}_0-(\mathbf{I}-\gamma\mathbf{H})^N\mathbf{B}_0(\mathbf{I}-\gamma\mathbf{H})^N)\right)=\sum_i\big(1-(1-\gamma\lambda_i)^{2N}\big)\cdot\big(\langle\mathbf{w}_0-\mathbf{w}^*,\mathbf{v}_i\rangle\big)^2.$$

Note that

$$\big(1-(1-\gamma\lambda_i)^{2N}\big)\leq\min\{2,2N\gamma\lambda_i\},$$

thus it follows that,

$$\mathrm{tr}(\mathbf{B}_{0,N})\leq 2\sum_i\min\{1,N\gamma\lambda_i\}\big(\langle\mathbf{w}_0-\mathbf{w}^*,\mathbf{v}_i\rangle\big)^2\leq 2\big(\|\mathbf{w}_0-\mathbf{w}^*\|_{\mathbf{I}_{0:k^*}}^2+N\gamma\|\mathbf{w}_0-\mathbf{w}^*\|_{\mathbf{H}_{k^*:\infty}}^2\big). \tag{B.22}$$

where $k^*=\max\{k:\lambda_k\geq\frac{1}{N\gamma}\}$. Then plug this bound into (B.21), we have

$$I_1\leq\frac{2\alpha\big(\|\mathbf{w}_0-\mathbf{w}^*\|_{\mathbf{I}_{0:k^*}}^2+N\gamma\|\mathbf{w}_0-\mathbf{w}^*\|_{\mathbf{H}_{k^*:\infty}}^2\big)}{N\gamma(1-\gamma\alpha\,\mathrm{tr}(\mathbf{H}))}\cdot\bigg(\frac{k^*}{N}+N\gamma^2\sum_{i>k^*}\lambda_i^2\bigg), \tag{B.23}$$

In the sequel we will upper bound I_2. Let $\mathbf{H}=\mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$ be the orthogonal decomposition of $\mathbf{H}$, where $\mathbf{V}=(\mathbf{v}_1,\mathbf{v}_2,\dots)$ and $\mathbf{\Lambda}$ is a diagonal matrix with diagonal entries $\lambda_1,\lambda_2,\dots$. Then we have

$$I_2=\frac{1}{\gamma N^2}\sum_{k=0}^{N-1}\big\langle(\mathbf{I}-\gamma\mathbf{\Lambda})^k-(\mathbf{I}-\gamma\mathbf{\Lambda})^{N+k},\mathbf{V}^\top\mathbf{B}_0\mathbf{V}\big\rangle.$$

Note that $(\mathbf{I}-\gamma\mathbf{\Lambda})^k-(\mathbf{I}-\gamma\mathbf{\Lambda})^{N+k}$ is a diagonal matrix, thus the above inner product only operates on the diagonal entries of $\mathbf{V}^\top\mathbf{B}_0\mathbf{V}$. Note that $\mathbf{B}_0=\boldsymbol{\eta}_0\boldsymbol{\eta}_0^\top$, it can be shown that the diagonal entries of $\mathbf{V}^\top\mathbf{B}_0\mathbf{V}$ are $\omega_1^2,\omega_2^2,\dots$, where $\omega_i=\mathbf{v}_i^\top\boldsymbol{\eta}_0=\mathbf{v}_i^\top(\mathbf{w}_0-\mathbf{w}^*)$.

$$\begin{aligned}
I_2&=\frac{1}{\gamma N^2}\sum_{k=0}^{N-1}\big\langle(\mathbf{I}-\gamma\mathbf{H})^k-(\mathbf{I}-\gamma\mathbf{H})^{N+k},\mathbf{B}_0\big\rangle\\
&=\frac{1}{\gamma N^2}\sum_{k=0}^{N-1}\sum_i\big[(1-\gamma\lambda_i)^k-(1-\gamma\lambda_i)^{N+k}\big]\omega_i^2\\
&=\frac{1}{\gamma^2N^2}\sum_i\frac{\omega_i^2}{\lambda_i}\big[1-(1-\gamma\lambda_i)^N\big]^2\\
&\leq\frac{1}{\gamma^2N^2}\sum_i\frac{\omega_i^2}{\lambda_i}\min\big\{1,\gamma^2N^2\lambda_i^2\big\}\\
&\leq\frac{1}{\gamma^2N^2}\cdot\sum_{i\leq k^*}\frac{\omega_i^2}{\lambda_i}+\sum_{i>k^*}\lambda_i\omega_i^2
\end{aligned}$$

$$= \frac{1}{\gamma^2 N^2} \cdot \|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k^*}} + \|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}},$$

where the first inequality is by (B.12) and $k^* = \max\{k : \lambda_k \geq \gamma^{-1}/N\}$. Combining the upper bounds on I_1 and I_2 directly completes the proof. ■

B.5 Proof of Theorem 2.1

Proof By Lemma B.2, it suffices to substitute into the upper bounds on the bias and variance errors. In particular, by Young's inequality we have

$$\mathbb{E}[L(\overline{\mathbf{w}}_N)] - L(\mathbf{w}^*) \leq \left(\sqrt{\mathrm{bias}} + \sqrt{\mathrm{variance}}\right)^2 \leq 2 \cdot \mathrm{bias} + 2 \cdot \mathrm{variance}.$$

Then we can directly substitute the bounds of variance and bias we proved in Lemmas B.6 and B.11. In particular, by Assumptions 2.2 we can directly get $R^2 = \alpha \operatorname{tr}(\mathbf{H})$. Therefore, it holds that

$$\begin{aligned}
&\mathbb{E}[L(\overline{\mathbf{w}}_N)] - L(\mathbf{w}^*) \\
&\leq 2\bigg[\frac{\alpha\|\mathbf{w}_0 - \mathbf{w}^*\|_2^2}{\gamma(1-\gamma\alpha\operatorname{tr}(\mathbf{H}))} \cdot \bigg(\frac{k^*}{N^2} + \gamma^2 \sum_{i>k^*} \lambda_i^2\bigg) + \frac{1}{\gamma^2 N^2} \cdot \|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k^*}} + \|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}} \\
&\qquad + \frac{\sigma_z^2}{1-\gamma\alpha\operatorname{tr}(\mathbf{H})}\bigg(\frac{k^*}{N} + \gamma^2 N \cdot \sum_{i>k^*} \lambda_i^2\bigg)\bigg] \\
&= 2 \cdot \mathrm{EffectiveBias} + 2 \cdot \mathrm{EffectiveVar},
\end{aligned}$$

where

$$\begin{aligned}
\mathrm{EffectiveBias} &= \frac{1}{\gamma^2 N^2} \cdot \|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k^*}} + \|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}} \\
\mathrm{EffectiveVar} &= \bigg(\frac{\sigma_z^2}{1-\gamma\alpha\operatorname{tr}(\mathbf{H})} + \frac{\alpha\|\mathbf{w}_0 - \mathbf{w}^*\|_2^2}{N\gamma(1-\gamma\alpha\operatorname{tr}(\mathbf{H}))}\bigg)\bigg(\frac{k^*}{N} + \gamma^2 N \cdot \sum_{i>k^*} \lambda_i^2\bigg).
\end{aligned}$$

■

B.6 Proof of Corollary 2.2

Proof We will show that the corollary can be directly implied by Theorem 2.1. In terms of the effective bias term, it is clear that

$$\begin{aligned}
\mathrm{EffectiveBias} &\leq \frac{1}{\gamma^2 N^2} \cdot \|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k^*}} + \|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}} \\
&= \frac{1}{\gamma^2 N^2} \cdot \lambda_{k^*}^{-1} \sum_{i \leq k^*} \left(\mathbf{v}_i^\top \mathbf{w}_0 - \mathbf{v}_i^\top \mathbf{w}^*\right)^2 + \lambda_{k^*+1} \sum_{i>k^*} \left(\mathbf{v}_i^\top \mathbf{w}_0 - \mathbf{v}_i^\top \mathbf{w}^*\right)^2.
\end{aligned}$$

where $\mathbf{v}_i$ is the eigenvector of $\mathbf{H}$ corresponding to the eigenvalue λ_i. Based on our definition of k^*, we have $\lambda_{k^*}^{-1} \leq N\gamma$ and $\lambda_{k^*+1} \leq 1/(N\gamma)$. Therefore, it follows that

$$\text{EffectiveBias} \leq \frac{1}{\gamma N} \cdot \sum_i \left(\mathbf{v}_i^\top \mathbf{w}_0 - \mathbf{v}_i^\top \mathbf{w}^*\right)^2 = \frac{\|\mathbf{w}_0 - \mathbf{w}^*\|_2^2}{\gamma N}. \tag{B.24}$$

Then regarding the effective variance, given the choice of stepsize that $\gamma = 1/(2\alpha \operatorname{tr}(\mathbf{H}))$, we have

$$\begin{aligned}\text{EffectiveVar} &\leq 2\bigg(\sigma^2 + \frac{\alpha\|\mathbf{w}_0 - \mathbf{w}^*\|_2^2}{N\gamma}\bigg)\left(\frac{k^*}{N} + \gamma^2 N \cdot \sum_{i>k^*} \lambda_i^2\right) \\ &= 2\sigma^2 \cdot \left(\frac{k^*}{N} + \gamma^2 N \cdot \sum_{i>k^*} \lambda_i^2\right) + \frac{2\alpha\|\mathbf{w}_0 - \mathbf{w}^*\|_2^2}{\gamma N} \cdot \left(\frac{k^*}{N} + \gamma^2 N \cdot \sum_{i>k^*} \lambda_i^2\right).\end{aligned}$$

Based on the definition of k^*, we have $\lambda_i \leq 1/(N\gamma)$ for $i > k^*$, thus

$$\gamma^2 N \sum_{i>k^*} \lambda_i^2 \leq \gamma \sum_{i>k^*} \lambda_i.$$

Besides, we also have $k^*/N \leq \gamma \sum_{i=1}^{k^*} \lambda_i$. Therefore, we have

$$\text{EffectiveVar} \leq 2\sigma^2 \cdot \left(\frac{k^*}{N} + \gamma^2 N \cdot \sum_{i>k^*} \lambda_i^2\right) + \frac{2\gamma\alpha\|\mathbf{w}_0 - \mathbf{w}^*\|_2^2}{\gamma N} \cdot \sum_i \lambda_i.$$

According to our choice of stepsize that $\gamma = 1/(2\alpha \operatorname{tr}(\mathbf{H}))$, we can get

$$\frac{2\gamma\alpha\|\mathbf{w}_0 - \mathbf{w}^*\|_2^2}{\gamma N} \cdot \sum_i \lambda_i = \frac{\|\mathbf{w}_0 - \mathbf{w}^*\|_2^2}{\gamma N}.$$

This further implies that

$$\text{EffectiveVar} \leq 2\sigma^2 \cdot \left(\frac{k^*}{N} + \gamma^2 N \cdot \sum_{i>k^*} \lambda_i^2\right) + \frac{\|\mathbf{w}_0 - \mathbf{w}^*\|_2^2}{\gamma N}. \tag{B.25}$$

Combining (B.24) and (B.25), we have

$$\begin{aligned}\mathbb{E}[L(\overline{\mathbf{w}}_N)] - L(\mathbf{w}^*) &\leq 2 \cdot \text{EffectiveBias} + 2 \cdot \text{EffectiveVar} \\ &\leq \frac{4\|\mathbf{w}_0 - \mathbf{w}^*\|_2^2}{\gamma N} + 4\sigma^2 \cdot \left(\frac{k^*}{N} + \gamma^2 N \cdot \sum_{i>k^*} \lambda_i^2\right).\end{aligned}$$

Further using the assumption that $\gamma = 1/(2\alpha \operatorname{tr}(\mathbf{H}))$ completes the proof. ■

B.7 Proof of Corollary 2.3

Proof For the bias error term, recall the definition of k^*, we have

$$\begin{aligned}\mathrm{EffectiveBias} &\le \mathcal{O}\left(\frac{1}{N^2}\cdot\|\mathbf{w}_0-\mathbf{w}^*\|_{\mathbf{H}_{0:k^*}^{-1}}^2+\|\mathbf{w}_0-\mathbf{w}^*\|_{\mathbf{H}_{k^*:\infty}}^2\right)\\ &\le \mathcal{O}\left(\frac{1}{N^2}\cdot\frac{1}{\lambda_{k^*}}\cdot\|\mathbf{w}_0-\mathbf{w}^*\|_2^2+\lambda_{k^*}\cdot\|\mathbf{w}_0-\mathbf{w}^*\|_2^2\right)\\ &\le \mathcal{O}\left(\frac{1}{N}\right).\end{aligned}$$

For the variance error term, it can be verified that all these examples satisfies $\sum_i \lambda_i<\infty$, thus we have

$$\mathrm{EffectiveVar}=\mathcal{O}\left(\frac{k^*}{N}+N\sum_{i>k^*}\lambda_i^2\right).$$

1. By the definition of k^* we have $k^*=s=N^r$, therefore

$$\mathrm{EffectiveVar}=\mathcal{O}\left(N^{-1}\cdot N^r+N\cdot N^{-q}\right)=\mathcal{O}\left(N^{r-1}+N^{1-q}\right).$$

2. By the definition of k^* we have $k^*=\Theta\big(N^{1/(1+r)}\big)$, therefore

$$\mathrm{EffectiveVar}=\mathcal{O}\left(N^{-1}\cdot N^{1/(1+r)}+N\cdot\left(N^{1/(1+r)}\right)^{-1-2r}\right)=\mathcal{O}\left(N^{-r/(1+r)}\right).$$

3. By the definition of k^* it can be shown that $k^*=\Omega\big(N/\log^\beta(N)\big)$ since otherwise

$$\lambda_{k^*+1}=\omega\bigg(\frac{\log^\beta(N)}{N}\cdot\frac{1}{\big[\log(N)-\beta\log(\log(N))\big]^\beta}\bigg)=\omega(1/N),$$

which contradicts to the fact that $\lambda_{k^*+1}=\mathcal{O}(1/N)$. Besides, we have

$$\sum_{i\ge k^*}\lambda_i^2=\mathcal{O}\left(\int_{k^*}^\infty\frac{1}{x^2\log^{2\beta}(x+1)}\mathrm{d}x\right).$$

Then note that

$$\frac{1}{x^2\log^{2\beta}(x+1)}\le\frac{\log^{2\beta}(x+1)+2\beta x\log^{2\beta-1}(x)/(x+1)}{x^2\log^{4\beta}(x)}.$$

This implies that

$$\begin{aligned}\int_{k^*}^\infty\frac{1}{x^2\log^{2\beta}(x)}\mathrm{d}x &\le \int_{k^*}^\infty\frac{\log^{2\beta}(x+1)+2\beta x\log^{2\beta-1}(x)/(x+1)}{x^2\log^{4\beta}(x)}\mathrm{d}x\\ &=\frac{1}{k^*\log^{2\beta}(k^*+1)}\end{aligned}$$

$$= \mathcal{O}\left(N^{-1}\log^{-\beta}(k^*)\right),$$

where the last equality is due to the fact that $1/(k^*\log^{\beta}(k^*+1)) = \Theta(1/N)$. As a result, we can get

$$\mathrm{EffectiveVar} = \mathcal{O}\left(k^*\cdot N^{-1} + N\sum_{i\geq k^*}\lambda_i^2\right) = \mathcal{O}\left(\log^{-\beta}(k^*)\right) = \mathcal{O}\left(\log^{-\beta}(N)\right),$$

where the second equality is due to the fact that $k*/N = \mathcal{O}\left(\log^{-\beta}(k^*)\right)$ and the last equality is due to $k^* = \Omega\big(N/\log^{\beta}(N)\big)$.

4. By definition of k^* we have $k^* = \Theta(\log N)$, therefore

$$\mathrm{EffectiveVar} = \mathcal{O}\left(N^{-1}\cdot\log N + N\cdot e^{-2\log N}\right) = \mathcal{O}\left(N^{-1}\log N\right).$$

Summing up the bias error and variance error concludes the proof. ■

Appendix C. Proofs of the Lower Bounds

C.1 Lower Bound for Bias-Variance Decomposition

We first introduce the following lemma to lower bound the excess risk when the noise is well-specified as in (2.1).

Lemma C.1 *Suppose the model noise ξ_t is well-specified, i.e., ξ_t and $\mathbf{x}_t$ are independent and $\mathbb{E}[\xi_t]=0$. Then*

$$\mathbb{E}[L(\overline{\mathbf{w}}_N) - L(\mathbf{w}^*)] \geq \frac{1}{2N^2}\cdot\sum_{t=0}^{N-1}\sum_{k=t}^{N-1}\left\langle(\mathbf{I}-\gamma\mathbf{H})^{k-t}\mathbf{H},\mathbf{B}_t\right\rangle + \frac{1}{2N^2}\cdot\sum_{t=0}^{N-1}\sum_{k=t}^{N-1}\left\langle(\mathbf{I}-\gamma\mathbf{H})^{k-t}\mathbf{H},\mathbf{C}_t\right\rangle.$$

Proof Let $\mathbf{P}_t = \mathbf{I}-\gamma\mathbf{x}_t\mathbf{x}_t^\top$, then the definitions of $\boldsymbol{\eta}_t^{\mathrm{bias}}$ in (4.3) and $\boldsymbol{\eta}_t^{\mathrm{variance}}$ (4.2) imply

$$\boldsymbol{\eta}_t^{\mathrm{bias}} = \prod_{k=1}^{t}\mathbf{P}_k\boldsymbol{\eta}_0, \qquad \boldsymbol{\eta}_t^{\mathrm{variance}} = \gamma\sum_{i=1}^{t}\prod_{j=i+1}^{t}\xi_i\mathbf{P}_j\mathbf{x}_i.$$

Note that in the well specified case, the noise $\xi_t := y_t - \langle\mathbf{w}^*,\mathbf{x}_t\rangle$ is independent of the data $\mathbf{x}_t$, and is of zero mean, hence

$$\mathbb{E}[\boldsymbol{\eta}_t^{\mathrm{bias}}\otimes\boldsymbol{\eta}_t^{\mathrm{variance}}] = \gamma\mathbb{E}\bigg[\prod_{k=1}^{t}\mathbf{P}_k\boldsymbol{\eta}_0\otimes\sum_{i=1}^{t}\prod_{j=i+1}^{t}\xi_i\mathbf{P}_j\mathbf{x}_i\bigg]$$

$$= \gamma \sum_{i=1}^{t} \mathbb{E}\big[\prod_{k=1}^{t} \mathbf{P}_k \boldsymbol{\eta}_0 \otimes \prod_{j=i+1}^{t} \mathbf{P}_j \mathbf{x}_i \big] \cdot \mathbb{E}[\xi_i] = \mathbf{0}.$$

This implies that

$$\mathbb{E}[\bar{\boldsymbol{\eta}}_t \otimes \bar{\boldsymbol{\eta}}_t] = \mathbb{E}[\bar{\boldsymbol{\eta}}_t^{\mathrm{bias}} \otimes \bar{\boldsymbol{\eta}}_t^{\mathrm{bias}}] + \mathbb{E}[\bar{\boldsymbol{\eta}}_t^{\mathrm{variance}} \otimes \bar{\boldsymbol{\eta}}_t^{\mathrm{variance}}],$$

and furthermore,

$$\begin{aligned} \mathbb{E}[L(\overline{\mathbf{w}}_N) - L(\mathbf{w}^*)] &= \frac{1}{2} \langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\eta}}_t \otimes \bar{\boldsymbol{\eta}}_t] \rangle \\ &= \frac{1}{2} \langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\eta}}_t^{\mathrm{bias}} \otimes \bar{\boldsymbol{\eta}}_t^{\mathrm{bias}}] \rangle + \frac{1}{2} \langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\eta}}_t^{\mathrm{variance}} \otimes \bar{\boldsymbol{\eta}}_t^{\mathrm{variance}}] \rangle. \end{aligned} \tag{C.1}$$

Next, we lower bound each term on the R.H.S. of (C.1) separately. By (B.5), we have

$$\mathbb{E}[\bar{\boldsymbol{\eta}}_N^{\mathrm{bias}} \otimes \bar{\boldsymbol{\eta}}_N^{\mathrm{bias}}] = \frac{1}{N^2} \cdot \bigg(\sum_{0 \le k < t \le N-1} \mathbb{E}[\boldsymbol{\eta}_t^{\mathrm{bias}} \otimes \boldsymbol{\eta}_k^{\mathrm{bias}}] + \sum_{0 \le t \le k \le N-1} \mathbb{E}[\boldsymbol{\eta}_t^{\mathrm{bias}} \otimes \boldsymbol{\eta}_k^{\mathrm{bias}}] \bigg).$$

Additionally, by (B.3) we can get

$$\begin{aligned} \bigg\langle \mathbf{H}, \sum_{0 \le k < t \le N-1} \mathbb{E}[\boldsymbol{\eta}_t^{\mathrm{bias}} \otimes \boldsymbol{\eta}_k^{\mathrm{bias}}] \bigg\rangle &= \bigg\langle \mathbf{H}, \sum_{k=0}^{N-1} \sum_{t=k+1}^{N-1} (\mathbf{I} - \gamma \mathbf{H})^{t-k} \mathbb{E}[\boldsymbol{\eta}_k^{\mathrm{bias}} \otimes \boldsymbol{\eta}_k^{\mathrm{bias}}] \bigg\rangle \\ &= \sum_{k=0}^{N-1} \sum_{t=k+1}^{N-1} \big\langle (\mathbf{I} - \gamma \mathbf{H})^{t-k} \mathbf{H}, \mathbb{E}[\boldsymbol{\eta}_k^{\mathrm{bias}} \otimes \boldsymbol{\eta}_k^{\mathrm{bias}}] \big\rangle \ge 0, \end{aligned}$$

where the inequality is due to the fact that $(\mathbf{I} - \gamma \mathbf{H})^{t-k} \mathbf{H}$ and $\mathbb{E}[\boldsymbol{\eta}_k^{\mathrm{bias}} \otimes \boldsymbol{\eta}_k^{\mathrm{bias}}]$ are both PSD. Therefore, it follows that

$$\begin{aligned} \mathrm{bias} :=& \frac{1}{2} \langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\eta}}_N^{\mathrm{bias}} \otimes \bar{\boldsymbol{\eta}}_N^{\mathrm{bias}}] \rangle \\ &\ge \frac{1}{2N^2} \cdot \bigg\langle \mathbf{H}, \sum_{0 \le t \le k \le N-1} \mathbb{E}[\boldsymbol{\eta}_t^{\mathrm{bias}} \otimes \boldsymbol{\eta}_k^{\mathrm{bias}}] \bigg\rangle \\ &= \frac{1}{2N^2} \cdot \sum_{t=0}^{N-1} \sum_{k=t}^{N-1} \Big\langle \mathbf{H}, \mathbb{E}[\boldsymbol{\eta}_t^{\mathrm{bias}} \otimes \boldsymbol{\eta}_t^{\mathrm{bias}}] \cdot (\mathbf{I} - \gamma \mathbf{H})^{k-t} \Big\rangle \\ &= \frac{1}{2N^2} \cdot \sum_{t=0}^{N-1} \sum_{k=t}^{N-1} \Big\langle (\mathbf{I} - \gamma \mathbf{H})^{k-t} \mathbf{H}, \mathbb{E}[\boldsymbol{\eta}_t^{\mathrm{bias}} \otimes \boldsymbol{\eta}_t^{\mathrm{bias}}] \Big\rangle, \end{aligned} \tag{C.2}$$

where the last equality holds since $\mathbf{H}$ and $(\mathbf{I} - \gamma \mathbf{H})^{k-t}$ commute. Repeating the computation for the variance terms, we can similarly obtain

$$\begin{aligned} \mathrm{variance} :=& \frac{1}{2} \langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\eta}}_N^{\mathrm{variance}} \otimes \bar{\boldsymbol{\eta}}_N^{\mathrm{variance}}] \rangle \\ &\ge \frac{1}{2N^2} \cdot \sum_{t=0}^{N-1} \sum_{k=t}^{N-1} \Big\langle (\mathbf{I} - \gamma \mathbf{H})^{k-t} \mathbf{H}, \mathbb{E}[\boldsymbol{\eta}_t^{\mathrm{variance}} \otimes \boldsymbol{\eta}_t^{\mathrm{variance}}] \Big\rangle. \end{aligned} \tag{C.3}$$

Plugging (C.2) and (C.3) into (C.1) gives

$$\begin{aligned}
\mathbb{E}[L(\overline{\mathbf{w}}_N) - L(\mathbf{w}^*)] &= \frac{1}{2}\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\eta}}_t \otimes \bar{\boldsymbol{\eta}}_t] \rangle = \text{bias} + \text{variance} \\
&\geq \frac{1}{2N^2} \cdot \sum_{t=0}^{N-1} \sum_{k=t}^{N-1} \Big\langle (\mathbf{I} - \gamma \mathbf{H})^{k-t} \mathbf{H}, \mathbb{E}[\boldsymbol{\eta}_t^{\text{bias}} \otimes \boldsymbol{\eta}_t^{\text{bias}}] \Big\rangle \\
&\quad + \frac{1}{2N^2} \cdot \sum_{t=0}^{N-1} \sum_{k=t}^{N-1} \Big\langle (\mathbf{I} - \gamma \mathbf{H})^{k-t} \mathbf{H}, \mathbb{E}[\boldsymbol{\eta}_t^{\text{variance}} \otimes \boldsymbol{\eta}_t^{\text{variance}}] \Big\rangle.
\end{aligned}$$

■

C.2 Lower Bounding the Variance Error

Lemma C.2 *Suppose Assumptions 2.1 hold. Suppose the noise is well-specified as in* (2.1). *If the stepsize satisfies $\gamma < 1/\lambda_1$, it holds that*

$$\mathbf{C}_t \succeq \frac{\gamma \sigma_{\text{noise}}^2}{2} \left(\mathbf{I} - (\mathbf{I} - \gamma \mathbf{H})^{2t} \right).$$

Proof Recall that $\mathcal{M} - \widetilde{\mathcal{M}}$ is a PSD mapping by Lemma 4.1 and $\mathbf{C}_{t-1}$ is PSD, then from (4.5) we have

$$\begin{aligned}
\mathbf{C}_t &= (\mathcal{I} - \gamma \mathcal{T}) \circ \mathbf{C}_{t-1} + \gamma^2 \boldsymbol{\Sigma} \\
&= (\mathcal{I} - \gamma \widetilde{\mathcal{T}}) \circ \mathbf{C}_{t-1} + (\mathcal{M} - \widetilde{\mathcal{M}}) \circ \mathbf{C}_{t-1} + \gamma^2 \boldsymbol{\Sigma} \\
&\succeq (\mathcal{I} - \gamma \widetilde{\mathcal{T}}) \circ \mathbf{C}_{t-1} + \gamma^2 \sigma_{\text{noise}}^2 \mathbf{H} \qquad \text{(since in the well-specified case } \boldsymbol{\Sigma} = \sigma_{\text{noise}}^2 \mathbf{H}) \\
&\succeq \gamma^2 \sigma_{\text{noise}}^2 \cdot \sum_{k=0}^{t-1} (\mathcal{I} - \gamma \widetilde{\mathcal{T}})^k \circ \mathbf{H} \qquad \text{(solving the recursion)} \\
&= \gamma^2 \sigma_{\text{noise}}^2 \cdot \sum_{k=0}^{t-1} (\mathbf{I} - \gamma \mathbf{H})^k \mathbf{H} (\mathbf{I} - \gamma \mathbf{H})^k \qquad \text{(by the property of } \mathcal{I} - \gamma \widetilde{\mathcal{T}} \text{ in (4.1))} \\
&= \gamma^2 \sigma_{\text{noise}}^2 \cdot \left(\mathbf{I} - (\mathbf{I} - \gamma \mathbf{H})^{2t} \right) \cdot \left(2\gamma \mathbf{I} - \gamma^2 \mathbf{H} \right)^{-1} \\
&\succeq \frac{\gamma \sigma_{\text{noise}}^2}{2} \cdot \left(\mathbf{I} - (\mathbf{I} - \gamma \mathbf{H})^{2t} \right),
\end{aligned}$$

where in the last inequality we use $2\gamma \mathbf{I} - \gamma^2 \mathbf{H} \preceq 2\gamma \mathbf{I}$. This completes the proof. ■

Lemma C.3 *Suppose Assumptions 2.1 hold. Suppose the noise is well-specified as in* (2.1) *and $N \geq 500$. Denote*

$$\text{variance} = \frac{1}{2N^2} \cdot \sum_{t=0}^{N-1} \sum_{k=t}^{N-1} \Big\langle (\mathbf{I} - \gamma \mathbf{H})^{k-t} \mathbf{H}, \mathbf{C}_t \Big\rangle.$$

If the stepsize satisfies $\gamma < 1/\lambda_1$, then

$$\text{variance} \geq \frac{\sigma_{\text{noise}}^2}{50}\left(\frac{k^*}{N} + N\gamma^2 \cdot \sum_{i>k^*}\lambda_i^2\right),$$

where $k^* = \max\{k : \lambda_k \geq \frac{1}{N\gamma}\}$.

Proof We can lower bound the variance error as follows

$$\begin{aligned}
\text{variance} &= \frac{1}{2N^2}\sum_{t=0}^{N-1}\sum_{k=t}^{N-1}\big\langle(\mathbf{I}-\gamma\mathbf{H})^{k-t}\mathbf{H}, \mathbf{C}_t\big\rangle \\
&= \frac{1}{2\gamma N^2}\sum_{t=0}^{N-1}\big\langle \mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N-t}, \mathbf{C}_t\big\rangle \\
&\geq \frac{\sigma_{\text{noise}}^2}{4N^2}\sum_{t=0}^{N-1}\big\langle \mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N-t}, \mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{2t}\big\rangle \qquad \text{(use Lemma C.2)} \\
&= \frac{\sigma_{\text{noise}}^2}{4N^2}\sum_{i}\sum_{t=0}^{N-1}\left(1-(1-\gamma\lambda_i)^{N-t}\right)\left(1-(1-\gamma\lambda_i)^{2t}\right) \\
&\geq \frac{\sigma_{\text{noise}}^2}{4N^2}\sum_{i}\sum_{t=0}^{N-1}\left(1-(1-\gamma\lambda_i)^{N-t-1}\right)\left(1-(1-\gamma\lambda_i)^{t}\right),
\end{aligned}$$

where $\{\lambda_i\}_{i\geq 1}$ are the eigenvalues of $\mathbf{H}$ and are sorted in decreasing order. Define

$$f(x) := \sum_{t=0}^{N-1}\left(1-(1-x)^{N-t-1}\right)\left(1-(1-x)^{t}\right), \qquad 0 < x < 1,$$

then

$$\text{variance} \geq \frac{\sigma_{\text{noise}}^2}{4N^2}\sum_{i\geq 1} f(\gamma\lambda_i).$$

Clearly $f(x)$ is increasing for $0 < x < 1$. Moreover:

$$\begin{aligned}
f(x) &= \sum_{t=0}^{N-1}\left(1-(1-x)^{N-1-t}-(1-x)^t+(1-x)^{N-1}\right) \\
&= N - 2\frac{1-(1-x)^N}{x} + N(1-x)^{N-1}.
\end{aligned}$$

Next we lower bound $f(x)$ within the range $\frac{1}{N} < x < 1$ and $0 < x < \frac{1}{N}$, respectively.

First consider $\frac{1}{N} \leq x < 1$. Notice that $f(x)$ is increasing and $\left(1-\frac{1}{N}\right)^N \geq \left(1-\frac{1}{500}\right)^{500} > 1.1/3$ if $N \geq 500$, thus for $\frac{1}{N} \leq x < 1$, we have

$$f(x) \geq N - 2N + 3N\cdot(1-1/N)^N \geq 0.1N.$$

On the other hand, note that we have the fourth-order derivative of $f(x)$ is positive when $x \in (0, 1/N)$, thus for $0 \le x \le 1/N$, we can perform third-order Taylor expansion on $f(x)$ at $x = 0$, which gives

$$\begin{aligned} f(x) &\ge \frac{N(N-1)(N-2)x^2}{6} - \frac{N(N-1)(N-2)(N-3)x^3}{12} \\ &\ge \frac{N(N-1)(N-2)x^2}{12} \qquad \text{(since } x \le 1/N\text{)} \\ &\ge \frac{2N^3x^2}{25}. \qquad \text{(since } N \ge 500\text{)} \end{aligned}$$

In sum,

$$f(x) \ge \begin{cases} \frac{N}{10}, & \frac{1}{N} \le x < 1, \\ \frac{2N^3}{25}x^2, & 0 < x < \frac{1}{N}. \end{cases}$$

Set $k^* = \max\{k : \lambda_k \ge \frac{1}{N\gamma}\}$, then

$$\begin{aligned} \text{variance} &\ge \frac{\sigma^2_{\text{noise}}}{4N^2} \sum_i f(\gamma\lambda_i) \\ &\ge \frac{\sigma^2_{\text{noise}}}{4N^2} \left(\frac{Nk^*}{10} + \frac{2N^3}{25}\gamma^2 \cdot \sum_{i>k^*} \lambda_i^2 \right) \\ &\ge \frac{\sigma^2_{\text{noise}}}{50} \left(\frac{k^*}{N} + N\gamma^2 \cdot \sum_{i>k^*} \lambda_i^2 \right). \end{aligned}$$

This completes the proof. ■

C.3 Lower Bounding the Bias Error

Recall that we have the following lower bound on the bias error

$$\text{bias} \ge \frac{1}{2N^2} \sum_{t=0}^{N-1} \sum_{k=t}^{N-1} \big\langle (\mathbf{I} - \gamma\mathbf{H})^{k-t}\mathbf{H}, \mathbf{B}_t \big\rangle,$$

from which we notice that

$$\begin{aligned} \text{bias} &\ge \frac{1}{2N^2} \sum_{t=0}^{N-1} \sum_{k=t}^{N-1} \big\langle (\mathbf{I} - \gamma\mathbf{H})^{k-t}\mathbf{H}, \mathbf{B}_t \big\rangle = \frac{1}{2\gamma N^2} \sum_{t=0}^{N-1} \big\langle \mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^{N-t}, \mathbf{B}_t \big\rangle \\ &\ge \frac{1}{2\gamma N^2} \sum_{t=0}^{N/2} \big\langle \mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^{N-t}, \mathbf{B}_t \big\rangle \\ &\ge \frac{1}{2\gamma N^2} \bigg\langle \mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^{N/2}, \sum_{t=0}^{N/2} \mathbf{B}_t \bigg\rangle. \end{aligned} \tag{C.4}$$

Let $\mathbf{S}_n := \sum_{t=0}^{n-1} \mathbf{B}_t$. Then the reminding challenge is to lower bound $\mathbf{S}_{N/2+1} = \sum_{t=0}^{N/2} \mathbf{B}_t$. Similarly to the idea of proving the upper bound, we first establish a crude lower bound on $\mathbf{S}_n$ then improve it to a fine lower bound.

Lemma C.4 *Suppose Assumptions 2.1 and 2.4 hold. If the stepsize satisfies $\gamma < 1/\lambda_1$, then for any $n \geq 2$, it holds that*

$$\mathbf{S}_n \succeq \frac{\beta}{4} \operatorname{tr}\left(\left(\mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^{n/2}\right)\mathbf{B}_0\right) \cdot \left(\mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^{n/2}\right) + \sum_{t=0}^{n-1} (\mathbf{I} - \gamma\mathbf{H})^t \cdot \mathbf{B}_0 \cdot (\mathbf{I} - \gamma\mathbf{H})^t.$$

Proof We first build a crude bound for $\mathbf{S}_n$. Recall that $\widetilde{\mathcal{T}} - \mathcal{T}$ is a PSD mapping by Lemma 4.1, then

$$\mathbf{S}_n = \sum_{t=0}^{n-1} \mathbf{B}_t = \sum_{t=0}^{n-1} (\mathcal{I} - \gamma\mathcal{T})^t \circ \mathbf{B}_0 \succeq \sum_{t=0}^{n-1} (\mathcal{I} - \gamma\widetilde{\mathcal{T}})^t \circ \mathbf{B}_0 = \sum_{t=0}^{n-1} (\mathbf{I} - \gamma\mathbf{H})^t \cdot \mathbf{B}_0 \cdot (\mathbf{I} - \gamma\mathbf{H})^t.$$

Now we apply Assumption 2.4 with the above crude bound to obtain that

$$\begin{aligned}
(\mathcal{M} - \widetilde{\mathcal{M}}) \circ \mathbf{S}_n &\succeq \beta \operatorname{tr}(\mathbf{H}\mathbf{S}_n)\mathbf{H} \\
&\succeq \beta \operatorname{tr}\left(\sum_{t=0}^{n-1} (\mathbf{I} - \gamma\mathbf{H})^{2t}\mathbf{H} \cdot \mathbf{B}_0\right)\mathbf{H} \\
&\succeq \beta \operatorname{tr}\left(\sum_{t=0}^{n-1} (\mathbf{I} - 2\gamma\mathbf{H})^{t}\mathbf{H} \cdot \mathbf{B}_0\right)\mathbf{H} \\
&= \frac{\beta}{2\gamma} \operatorname{tr}\left(\left(\mathbf{I} - (\mathbf{I} - 2\gamma\mathbf{H})^n\right)\mathbf{B}_0\right)\mathbf{H} \\
&\succeq \frac{\beta}{2\gamma} \operatorname{tr}\left(\left(\mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^n\right)\mathbf{B}_0\right)\mathbf{H}.
\end{aligned}$$

Next we use the above inequality to build a fine lower bound for $\mathbf{S}_n$:

$$\begin{aligned}
\mathbf{S}_n &= (\mathcal{I} - \gamma\mathcal{T}) \circ \mathbf{S}_{n-1} + \mathbf{B}_0 = (\mathcal{I} - \gamma\widetilde{\mathcal{T}}) \circ \mathbf{S}_{n-1} + \gamma^2(\mathcal{M} - \widetilde{\mathcal{M}}) \circ \mathbf{S}_{n-1} + \mathbf{B}_0 \\
&\succeq (\mathcal{I} - \gamma\widetilde{\mathcal{T}}) \circ \mathbf{S}_{n-1} + \frac{\beta\gamma}{2} \operatorname{tr}\left(\left(\mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^{n-1}\right)\mathbf{B}_0\right)\mathbf{H} + \mathbf{B}_0.
\end{aligned}$$

Solving the recursion we obtain

$$\begin{aligned}
\mathbf{S}_n &\succeq \sum_{t=0}^{n-1} (\mathcal{I} - \gamma\widetilde{\mathcal{T}})^t \circ \left\{\frac{\beta\gamma}{2} \operatorname{tr}\left(\left(\mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^{n-1-t}\right)\mathbf{B}_0\right)\mathbf{H} + \mathbf{B}_0\right\} \\
&= \frac{\beta\gamma}{2} \sum_{t=0}^{n-1} \operatorname{tr}\left(\left(\mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^{n-1-t}\right)\mathbf{B}_0\right) \cdot (\mathbf{I} - \gamma\mathbf{H})^{2t}\mathbf{H} \\
&\quad + \sum_{t=0}^{n-1} (\mathbf{I} - \gamma\mathbf{H})^t \cdot \mathbf{B}_0 \cdot (\mathbf{I} - \gamma\mathbf{H})^t.
\end{aligned}$$

For the first term, noticing the following:

$$
\begin{aligned}
&\sum_{t=0}^{n-1} \operatorname{tr}\left(\left(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{n-1-t}\right)\mathbf{B}_0\right)\cdot(\mathbf{I}-\gamma\mathbf{H})^{2t}\mathbf{H} \\
&\succeq \sum_{t=0}^{n-1} \operatorname{tr}\left(\left(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{n-1-t}\right)\mathbf{B}_0\right)\cdot(\mathbf{I}-2\gamma\mathbf{H})^{t}\mathbf{H} \\
&\succeq \sum_{t=0}^{n/2-1} \operatorname{tr}\left(\left(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{n-1-t}\right)\mathbf{B}_0\right)\cdot(\mathbf{I}-2\gamma\mathbf{H})^{t}\mathbf{H} \\
&\succeq \operatorname{tr}\left(\left(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{n/2}\right)\mathbf{B}_0\right)\cdot\sum_{t=0}^{n/2-1}(\mathbf{I}-2\gamma\mathbf{H})^{t}\mathbf{H} \\
&= \frac{1}{2\gamma}\operatorname{tr}\left(\left(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{n/2}\right)\mathbf{B}_0\right)\cdot\left(\mathbf{I}-(\mathbf{I}-2\gamma\mathbf{H})^{n/2}\right) \\
&\succeq \frac{1}{2\gamma}\operatorname{tr}\left(\left(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{n/2}\right)\mathbf{B}_0\right)\cdot\left(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{n/2}\right),
\end{aligned}
$$

inserting which back to the lower bound for $\mathbf{S}_n$, we complete the proof. ∎

Lemma C.5 *Suppose Assumptions 2.1 and 2.4 hold and $N \geq 2$. If the stepsize satisfies $\gamma < 1/\gamma_1$, then*

$$
\begin{aligned}
\mathrm{bias} \geq\ & \frac{1}{100\gamma^2N^2}\cdot\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k^*}} + \frac{1}{100}\cdot\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}} \\
&+ \frac{\beta\left(\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{I}_{0:k^*}}+\gamma N\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}}\right)}{1000\gamma N^2}\cdot\left(k^*+\gamma^2N^2\sum_{i>k^*}\lambda_i^2\right),
\end{aligned}
$$

where $k^* = \max\{k : \lambda_k \geq \frac{1}{N\gamma}\}$.

Proof According to (C.4) and Lemma C.4, we have that

$$
\begin{aligned}
\mathrm{bias} &\geq \frac{1}{2\gamma N^2}\left\langle \mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N/2}, \mathbf{S}_{N/2+1}\right\rangle \geq \frac{1}{2\gamma N^2}\left\langle \mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N/2}, \mathbf{S}_{N/2}\right\rangle \\
&\geq \underbrace{\frac{\beta}{8\gamma N^2}\operatorname{tr}\left(\left(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N/4}\right)\mathbf{B}_0\right)\cdot\left\langle \mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N/2}, \mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N/4}\right\rangle}_{I_1} \\
&\quad + \underbrace{\frac{1}{2\gamma N^2}\left\langle \mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N/2}, \sum_{t=0}^{N/2-1}(\mathbf{I}-\gamma\mathbf{H})^{t}\cdot\mathbf{B}_0\cdot(\mathbf{I}-\gamma\mathbf{H})^{t}\right\rangle}_{I_2}.
\end{aligned}
$$

The first term is lower bounded by

$$
I_1 \geq \frac{\beta}{8\gamma N^2}\operatorname{tr}\left(\left(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N/4}\right)\mathbf{B}_0\right)\cdot\operatorname{tr}\left(\left(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N/4}\right)^2\right)
$$

$$= \frac{\beta}{8\gamma N^2}\left(\sum_i \left(1-(1-\gamma\lambda_i)^{N/4}\right)\omega_i^2\right)\cdot\left(\sum_i \left(1-(1-\gamma\lambda_i)^{N/4}\right)^2\right),$$

where $\omega_i = \mathbf{v}_i^\top(\mathbf{w}_0-\mathbf{w}^*)$ for $\mathbf{v}_1,\dots,\mathbf{v}_d$ being the eigenvectors of $\mathbf{H}$; and the second term is lower bounded by

$$\begin{aligned}
I_2 &= \frac{1}{2\gamma N^2}\left\langle \sum_{t=0}^{N/2-1}(\mathbf{I}-\gamma\mathbf{H})^{2t}\left(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N/2}\right), \mathbf{B}_0\right\rangle \\
&\ge \frac{1}{2\gamma N^2}\left\langle \sum_{t=0}^{N/2-1}(\mathbf{I}-2\gamma\mathbf{H})^{t}\left(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N/2}\right), \mathbf{B}_0\right\rangle \\
&\ge \frac{1}{4\gamma^2 N^2}\left\langle \left(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N/2}\right)^2\mathbf{H}^{-1}, \mathbf{B}_0\right\rangle \\
&\ge \frac{1}{4\gamma^2 N^2}\left\langle \left(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N/4}\right)^2\mathbf{H}^{-1}, \mathbf{B}_0\right\rangle \\
&= \frac{1}{4\gamma^2 N^2}\sum_i\left(1-(1-\gamma\lambda_i)^{N/4}\right)^2\lambda_i^{-1}\omega_i^2.
\end{aligned}$$

To further lower bound the two terms, noticing the following inequality:

$$1-(1-\gamma\lambda_i)^{\frac{N}{4}} \ge \begin{cases} 1-(1-\frac{1}{N})^{\frac{N}{4}} \ge 1-e^{-\frac{1}{4}} \ge \frac{1}{5}, & \lambda_i \ge \frac{1}{\gamma N}, \\ \frac{N}{4}\cdot\gamma\lambda_i - \frac{N(N-4)}{32}\cdot\gamma^2\lambda_i^2 \ge \frac{N}{5}\cdot\gamma\lambda_i, & \lambda_i < \frac{1}{\gamma N}. \end{cases}$$

Plugging this into the bounds for I_1 and I_2, and setting $k^* := \max\{k : \lambda_k \ge 1/(\gamma N)\}$, we then obtain that

$$\begin{aligned}
I_1 &\ge \frac{\beta}{8\gamma N^2}\cdot\left(\frac{1}{5}\cdot\sum_{i\le k^*}\omega_i^2 + \frac{\gamma N}{5}\sum_{i>k^*}\lambda_i\omega_i^2\right)\cdot\left(\frac{1}{25}\cdot k^* + \frac{\gamma^2N^2}{25}\cdot\sum_{i>k^*}\lambda_i^2\right) \\
&= \frac{\beta}{1000\gamma N^2}\cdot\left(\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{I}_{0:k^*}} + \gamma N\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}}\right)\cdot\left(k^* + \gamma^2N^2\sum_{i>k^*}\lambda_i^2\right),
\end{aligned}$$

and that

$$\begin{aligned}
I_2 &\ge \frac{1}{4\gamma^2N^2}\left(\frac{1}{25}\cdot\sum_{i\le k^*}\lambda_i^{-1}\omega_i^2 + \frac{\gamma^2N^2}{25}\cdot\sum_{i>k^*}\lambda_i\omega_i^2\right) \\
&= \frac{1}{100\gamma^2N^2}\left(\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k^*}} + \gamma^2N^2\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}}\right).
\end{aligned}$$

Summing up the two terms completes the proof.

■

C.4 Proof of Theorem 2.2

Proof Plugging the bounds of the bias error and variance error in Lemmas C.5 and C.3 into Lemma C.1 immediately completes the proof. ■

Appendix D. Proofs for Tail-Averaging

In this section, we provide the proofs for SGD with tail-averaging. Recall that in tail-averaging, we take average from the s-th iterate, i.e., the output of the tail-average SGD is

$$\overline{\mathbf{w}}_{s:s+N} = \frac{1}{N}\sum_{t=s}^{s+N-1} \mathbf{w}_t.$$

D.1 Upper Bounds for Tail-Averaging

The following two lemmas are straightforward extensions of Lemmas B.2 and B.3.

Lemma D.1 (Variant of Lemma B.2)

$$\mathbb{E}[L(\overline{\mathbf{w}}_{s:s+N})] - L(\mathbf{w}^*) = \frac{1}{2}\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\eta}}_{s:s+N} \otimes \bar{\boldsymbol{\eta}}_{s:s+N}]\rangle \le \left(\sqrt{\mathrm{bias}} + \sqrt{\mathrm{variance}}\right)^2,$$

where

$$\mathrm{bias} := \frac{1}{2}\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\eta}}^{\mathrm{bias}}_{s:s+N} \otimes \bar{\boldsymbol{\eta}}^{\mathrm{bias}}_{s:s+N}]\rangle, \qquad \mathrm{variance} := \frac{1}{2}\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\eta}}^{\mathrm{variance}}_{s:s+N} \otimes \bar{\boldsymbol{\eta}}^{\mathrm{variance}}_{s:s+N}]\rangle.$$

Lemma D.2 (Variant of Lemma B.3) *Recall iterates* (4.4) *and* (4.5). *If the stepsize satisfies $\gamma < 1/\lambda_1$, the bias error and variance error are upper bounded respectively as follows:*

$$\mathrm{bias} := \frac{1}{2}\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\eta}}^{\mathrm{bias}}_{s:s+N} \otimes \bar{\boldsymbol{\eta}}^{\mathrm{bias}}_{s:s+N}]\rangle \le \frac{1}{N^2}\sum_{t=0}^{N-1}\sum_{k=t}^{N-1} \big\langle (\mathbf{I}-\gamma\mathbf{H})^{k-t}\mathbf{H}, \mathbf{B}_{s+t}\big\rangle,$$

$$\mathrm{variance} := \frac{1}{2}\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\eta}}^{\mathrm{variance}}_{s:s+N} \otimes \bar{\boldsymbol{\eta}}^{\mathrm{variance}}_{s:s+N}] \le \frac{1}{N^2}\sum_{t=0}^{N-1}\sum_{k=t}^{N-1} \big\langle (\mathbf{I}-\gamma\mathbf{H})^{k-t}\mathbf{H}, \mathbf{C}_{s+t}\big\rangle.$$

Proof By replacing $\mathbf{B}_0$ and $\mathbf{C}_0$ by $\mathbf{B}_s$ and $\mathbf{C}_s$ in the proof of Lemma B.3, and repeating the remaining arguments, we can easily complete the proof. ■

D.1.1 Bounding the Variance Error

Lemma D.3 (Variant of Lemma B.6) *Under Assumptions 2.1, 2.3 and B.1, if the stepsize satisfies $\gamma < 1/R^2$, then it holds that*

$$\mathrm{variance} \le \frac{\sigma^2}{1-\gamma R^2}\cdot\left(\frac{k^*}{N} + \gamma\cdot\sum_{k^*<i\le k^\dagger}\lambda_i + \gamma^2(s+N)\cdot\sum_{i>k^\dagger}\lambda_i^2\right),$$

where $k^* = \min\{k : \lambda_i < \frac{1}{\gamma N}\}$ *and* $k^\dagger = \min\{k : \lambda_i < \frac{1}{\gamma(s+N)}\}$.

Proof By Lemma D.2, we can bound the variance error as follows

$$\begin{aligned}
\text{variance} &\le \frac{1}{N^2}\sum_{t=0}^{N-1}\sum_{k=t}^{N-1}\big\langle (\mathbf{I}-\gamma\mathbf{H})^{k-t}\mathbf{H}, \mathbf{C}_{s+t}\big\rangle \\
&= \frac{1}{\gamma N^2}\sum_{t=0}^{N-1}\big\langle \mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N-t}, \mathbf{C}_{s+t}\big\rangle \\
&\le \frac{\sigma^2}{N^2(1-\gamma R^2)}\sum_{t=0}^{N-1}\big\langle \mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N-t}, \big(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{s+t}\big)\big\rangle \\
&= \frac{\sigma^2}{N^2(1-\gamma R^2)}\sum_i\sum_{t=0}^{N-1}\big(1-(1-\gamma\lambda_i)^{N-t}\big)\big(1-(1-\gamma\lambda_i)^{s+t}\big) \\
&\le \frac{\sigma^2}{N^2(1-\gamma R^2)}\sum_i\sum_{t=0}^{N-1}\big(1-(1-\gamma\lambda_i)^{N}\big)\big(1-(1-\gamma\lambda_i)^{s+N}\big) \\
&= \frac{\sigma^2}{N(1-\gamma R^2)}\sum_i\big(1-(1-\gamma\lambda_i)^{N}\big)\big(1-(1-\gamma\lambda_i)^{s+N}\big),
\end{aligned}$$

where the second inequality is due to Lemma B.5, $\{\lambda_i\}_{i\ge1}$ are the eigenvalues of $\mathbf{H}$ and are sorted in decreasing order. Now we will move to upper bound the quantity $\big(1-(1-\gamma\lambda_i)^{N}\big)\big(1-(1-\gamma\lambda_i)^{s+N}\big)$, which will be separately discussed according to the following three cases: (1) $\gamma\lambda_i \ge 1/N$, (2) $1/(s+N) \le \gamma\lambda_i < 1/N$, and (3) $\gamma\lambda < 1/(s+N)$. In case (1), we can crudely bound this quantity as follows,

$$\big(1-(1-\gamma\lambda_i)^{N}\big)\big(1-(1-\gamma\lambda_i)^{s+N}\big) \le 1.$$

In case (2), we can use $(1-\gamma\lambda_i)^N \ge 1-\gamma N\lambda_i$ and get

$$\big(1-(1-\gamma\lambda_i)^{N}\big)\big(1-(1-\gamma\lambda_i)^{s+N}\big) \le \gamma N\lambda_i\cdot 1 = \gamma N\lambda_i.$$

In case (3), we can use $(1-\gamma\lambda_i)^N \ge 1-\gamma N\lambda_i$ and $(1-\gamma\lambda_i)^{s+N} \ge 1-\gamma(s+N)\lambda_i$, and get

$$\big(1-(1-\gamma\lambda_i)^{N}\big)\big(1-(1-\gamma\lambda_i)^{s+N}\big) \le \gamma N\lambda_i\cdot\gamma(s+N)\lambda_i = \gamma^2N(s+N)\lambda_i^2.$$

Therefore, set $k^* = \min\{k : \lambda_i < \frac{1}{N\gamma}\}$ and $k^\dagger = \min\{k : \lambda_i < \frac{1}{(s+N)\gamma}\}$, we have

$$\begin{aligned}
\text{variance} &\le \frac{\sigma^2}{N(1-\gamma R^2)}\cdot\bigg(k^* + \gamma N\sum_{k^*<i\le k^\dagger}\lambda_i + \gamma^2N(s+N)\sum_{i>k^\dagger}\lambda_i^2\bigg) \\
&= \frac{\sigma^2}{1-\gamma R^2}\cdot\bigg(\frac{k^*}{N} + \gamma\cdot\sum_{k^*<i\le k^\dagger}\lambda_i + \gamma^2(s+N)\cdot\sum_{i>k^\dagger}\lambda_i^2\bigg).
\end{aligned}$$

This completes the proof. ∎

D.1.2 Bounding the Bias Error

Similarly to (B.13) and using Lemma D.2, we have the following upper bound for the bias error:

$$\begin{aligned}\mathrm{bias} &\le \frac{1}{N^2}\sum_{t=0}^{N-1}\sum_{k=t}^{N-1}\big\langle(\mathbf{I}-\gamma\mathbf{H})^{k-t}\mathbf{H},\mathbf{B}_{s+t}\big\rangle\\ &= \frac{1}{\gamma N^2}\sum_{t=0}^{N-1}\big\langle\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N-t},\mathbf{B}_{s+t}\big\rangle\\ &\le \frac{1}{\gamma N^2}\left\langle\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N},\sum_{t=0}^{N-1}\mathbf{B}_{s+t}\right\rangle. \qquad (\text{D.1})\end{aligned}$$

Let $\mathbf{S}_{s:s+t}=\sum_{k=s}^{s+t-1}\mathbf{B}_k$, then we only need to establish an upper bound for $\mathbf{S}_{s:s+N}$.

Lemma D.4 (Variant of Lemma B.10) *Let $\mathbf{S}_{s:s+t}=\sum_{k=s}^{s+t-1}\mathbf{B}_k$ for any $t\ge s$ and $\mathbf{B}_{a,b}=\mathbf{B}_a-(\mathbf{I}-\gamma\mathbf{H})^{b-a}\mathbf{B}_a(\mathbf{I}-\gamma\mathbf{H})^{b-a}$. Under Assumptions 2.1 and 2.2, if the stepsize satisfies $\gamma<1/\big(\alpha\operatorname{tr}(\mathbf{H})\big)$, it holds that*

$$\mathbf{S}_{s:s+N}\preceq\sum_{k=0}^{N-1}(\mathbf{I}-\gamma\mathbf{H})^{k+s}\mathbf{B}_0(\mathbf{I}-\gamma\mathbf{H})^{k+s}+\frac{\gamma\alpha\operatorname{tr}(\mathbf{B}_{s,s+N}+\mathbf{B}_{0,s})}{1-\gamma\alpha\operatorname{tr}(\mathbf{H})}\sum_{k=0}^{N-1}(\mathbf{I}-\gamma\mathbf{H})^{2k}\mathbf{H}.$$

Proof Based on the definition of $\mathbf{S}_{s:s+t}$, we have

$$\mathbf{S}_{s:s+t}=\sum_{k=s}^{s+t-1}\mathbf{B}_k=\sum_{k=0}^{t-1}(\mathcal{I}-\gamma\mathcal{T})^k\circ\mathbf{B}_s=(\mathcal{I}-\gamma\mathcal{T})\circ\mathbf{S}_{s:s+t-1}+\mathbf{B}_s.$$

Therefore, following the similar proof technique of Lemma B.10, we can get

$$\mathbf{S}_{s:s+N}\preceq\underbrace{\sum_{k=0}^{N-1}(\mathbf{I}-\gamma\mathbf{H})^{k}\mathbf{B}_s(\mathbf{I}-\gamma\mathbf{H})^{k}}_{I_1}+\underbrace{\frac{\gamma\alpha\operatorname{tr}(\mathbf{B}_{s,s+N})}{1-\gamma\alpha\operatorname{tr}(\mathbf{H})}\sum_{k=0}^{N-1}(\mathbf{I}-\gamma\mathbf{H})^{2k}\mathbf{H}}_{I_2}. \qquad (\text{D.2})$$

Now we will upper bound I_1, which requires a carefully characterization on $\mathbf{B}_s$. Particularly, the update form of $\mathbf{B}_k$ in (4.2) implies

$$\mathbf{B}_k=(\mathcal{I}-\gamma\mathcal{T})\circ\mathbf{B}_{k-1}\preceq(\mathcal{I}-\gamma\widetilde{\mathcal{T}})\circ\mathbf{B}_{k-1}+\gamma^2\mathcal{M}\circ\mathbf{B}_{k-1}.$$

By Assumption 2.2, we have $\mathcal{M}\circ\mathbf{B}_k\preceq\alpha\operatorname{tr}(\mathbf{H}\mathbf{B}_k)\cdot\mathbf{H}$. Thus,

$$\begin{aligned}\mathbf{B}_k &\preceq(\mathcal{I}-\gamma\widetilde{\mathcal{T}})\circ\mathbf{B}_{k-1}+\gamma^2\mathcal{M}\circ\mathbf{B}_{k-1}\\ &\preceq(\mathcal{I}-\gamma\widetilde{\mathcal{T}})\circ\mathbf{B}_{k-1}+\alpha\gamma^2\operatorname{tr}(\mathbf{H}\mathbf{B}_{k-1})\cdot\mathbf{H}\\ &=(\mathcal{I}-\gamma\widetilde{\mathcal{T}})^k\circ\mathbf{B}_0+\alpha\gamma^2\sum_{t=0}^{k-1}\operatorname{tr}(\mathbf{H}\mathbf{B}_t)\cdot(\mathcal{I}-\gamma\widetilde{\mathcal{T}})^{k-1-t}\circ\mathbf{H}\end{aligned}$$

$$\preceq (\mathcal{I} - \gamma\widetilde{\mathcal{T}})^k \circ \mathbf{B}_0 + \alpha\gamma^2 \sum_{t=0}^{k-1} \mathrm{tr}(\mathbf{H}\mathbf{B}_t) \cdot \mathbf{H} \tag{D.3}$$

where in the third inequality we use the fact that $\mathcal{I} - \gamma\widetilde{\mathcal{T}}$ is a PSD mapping and the last inequality is due to $(\mathcal{I} - \gamma\widetilde{\mathcal{T}})^{k-1-t}\mathbf{H} = (\mathbf{I} - \gamma\mathbf{H})^{2(k-1-t)}\mathbf{H} \preceq \mathbf{H}$. Next we will upper bound $\sum_{t=0}^{k-1} \mathrm{tr}(\mathbf{H}\mathbf{B}_t)$. Recall the definition of $\boldsymbol{\eta}_k^{\mathrm{bias}}$ and its update rule, we have

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{\eta}_k^{\mathrm{bias}}\|_2^2 | \boldsymbol{\eta}_{k-1}^{\mathrm{bias}}] &= \mathbb{E}[\|(\mathbf{I} - \gamma\mathbf{x}_k\mathbf{x}_k^\top)\boldsymbol{\eta}_{k-1}^{\mathrm{bias}}\|_2^2 | \boldsymbol{\eta}_{k-1}^{\mathrm{bias}}] \\
&= \|\boldsymbol{\eta}_{k-1}^{\mathrm{bias}}\|_2^2 - 2\gamma\mathbb{E}[\langle \mathbf{x}_k\mathbf{x}_k^\top, \boldsymbol{\eta}_{k-1}^{\mathrm{bias}} \otimes \boldsymbol{\eta}_{k-1}^{\mathrm{bias}}\rangle | \boldsymbol{\eta}_{k-1}^{\mathrm{bias}}] \\
&\quad + \gamma^2\mathbb{E}[\langle \mathbf{x}_k\mathbf{x}_k^\top\mathbf{x}_k\mathbf{x}_k^\top, \boldsymbol{\eta}_{k-1}^{\mathrm{bias}} \otimes \boldsymbol{\eta}_{k-1}^{\mathrm{bias}} | \boldsymbol{\eta}_{k-1}^{\mathrm{bias}}] \\
&= \|\boldsymbol{\eta}_{k-1}^{\mathrm{bias}}\|_2^2 - 2\gamma\langle \mathbf{H}, \boldsymbol{\eta}_{k-1}^{\mathrm{bias}} \otimes \boldsymbol{\eta}_{k-1}^{\mathrm{bias}}\rangle + \gamma^2\langle \mathcal{M} \circ \mathbf{I}, \boldsymbol{\eta}_{k-1}^{\mathrm{bias}} \otimes \boldsymbol{\eta}_{k-1}^{\mathrm{bias}}\rangle \\
&\le \|\boldsymbol{\eta}_{k-1}^{\mathrm{bias}}\|_2^2 - \big(2\gamma - \gamma^2\alpha\,\mathrm{tr}(\mathbf{H})\big) \cdot \langle \mathbf{H}, \boldsymbol{\eta}_{k-1}^{\mathrm{bias}} \otimes \boldsymbol{\eta}_{k-1}^{\mathrm{bias}}\rangle,
\end{aligned}$$

where the inequality is due to the fact that $\mathcal{M} \circ \mathbf{I} \preceq \alpha\,\mathrm{tr}(\mathbf{H})\mathbf{H}$. Note that $\mathbf{B}_k = \mathbb{E}[\boldsymbol{\eta}_k^{\mathrm{bias}} \otimes \boldsymbol{\eta}_k^{\mathrm{bias}}]$, taking total expectation further gives

$$\mathrm{tr}(\mathbf{B}_k) \le \mathrm{tr}(\mathbf{B}_{k-1}) - \big(2\gamma - \gamma^2\alpha\,\mathrm{tr}(\mathbf{H})\big) \cdot \mathrm{tr}(\mathbf{H}\mathbf{B}_{k-1}),$$

which implies that

$$\sum_{t=0}^{k-1} \mathrm{tr}(\mathbf{H}\mathbf{B}_t) \le \frac{\mathrm{tr}(\mathbf{B}_0) - \mathrm{tr}(\mathbf{B}_k)}{2\gamma - \gamma^2\alpha\,\mathrm{tr}(\mathbf{H})}. \tag{D.4}$$

Substituting (D.4) into (D.3) gives

$$\begin{aligned}
\mathbf{B}_k &\preceq (\mathcal{I} - \gamma\widetilde{\mathcal{T}})^k \circ \mathbf{B}_0 + \alpha\gamma^2 \sum_{t=0}^{k-1} \mathrm{tr}(\mathbf{H}\mathbf{B}_t) \cdot \mathbf{H} \\
&\preceq (\mathcal{I} - \gamma\widetilde{\mathcal{T}})^k \circ \mathbf{B}_0 + \frac{\gamma\alpha\,\mathrm{tr}(\mathbf{B}_0 - \mathbf{B}_k)}{2 - \gamma\alpha\,\mathrm{tr}(\mathbf{H})} \cdot \mathbf{H}.
\end{aligned}$$

Therefore, we further have

$$I_1 \preceq \sum_{k=0}^{N-1} (\mathbf{I} - \gamma\mathbf{H})^{k+s}\mathbf{B}_0(\mathbf{I} - \gamma\mathbf{H})^{k+s} + \frac{\gamma\alpha\,\mathrm{tr}(\mathbf{B}_0 - \mathbf{B}_s)}{2 - \gamma\alpha\,\mathrm{tr}(\mathbf{H})} \sum_{k=0}^{N-1} (\mathbf{I} - \gamma\mathbf{H})^{2k}\mathbf{H}. \tag{D.5}$$

Further note that $\mathbf{B}_s = (\mathcal{I} - \gamma\mathcal{T})^s\mathbf{B}_0$ and $\mathcal{T} \succeq \widetilde{\mathcal{T}}$, we have

$$\begin{aligned}
\mathrm{tr}(\mathbf{B}_0 - \mathbf{B}_s) &= \mathrm{tr}\left(\mathbf{B}_0 - (\mathcal{I} - \gamma\mathcal{T})^s\mathbf{B}_0\right) \\
&\le \mathrm{tr}\left(\mathbf{B}_0 - (\mathcal{I} - \gamma\widetilde{\mathcal{T}})^s\mathbf{B}_0\right) \\
&\le \mathrm{tr}\left(\mathbf{B}_0 - (\mathbf{I} - \gamma\mathbf{H})^s\mathbf{B}_0(\mathbf{I} - \gamma\mathbf{H})^s\right) \\
&= \mathrm{tr}(\mathbf{B}_{0,s}).
\end{aligned}$$

Now, we can substitute the above inequality and (D.5) into (D.2) and obtain the following upper bound on $\mathbf{S}_{s:s+N}$,

$$\mathbf{S}_{s:s+N} \preceq I_1 + I_2 \preceq \sum_{k=0}^{N-1} (\mathbf{I} - \gamma\mathbf{H})^{k+s}\mathbf{B}_0(\mathbf{I} - \gamma\mathbf{H})^{k+s} + \frac{\gamma\alpha\,\mathrm{tr}(\mathbf{B}_{s,s+N} + \mathbf{B}_{0,s})}{1 - \gamma\alpha\,\mathrm{tr}(\mathbf{H})} \sum_{k=0}^{N-1} (\mathbf{I} - \gamma\mathbf{H})^{2k}\mathbf{H},$$

where we use the fact that $0 \le 1 - \gamma\alpha\,\mathrm{tr}(\mathbf{H}) \le 2 - \gamma\alpha\,\mathrm{tr}(\mathbf{H})$. This completes the proof. ■

Lemma D.5 (Variant of Lemma B.11) *Under Assumptions 2.1 and 2.2, if the stepsize satisfies $\gamma < 1/(\alpha\,\mathrm{tr}(\mathbf{H}))$, it holds that*

$$\begin{aligned}\mathrm{bias} \le \frac{1}{\gamma^2 N^2} \cdot \big\|(\mathbf{I}-\gamma\mathbf{H})^s(\mathbf{w}_0-\mathbf{w}^*)\big\|^2_{\mathbf{H}^{-1}_{0:k^*}} + \big\|(\mathbf{I}-\gamma\mathbf{H})^s(\mathbf{w}_0-\mathbf{w}^*)\big\|^2_{\mathbf{H}_{k^*:\infty}} \\ + \frac{4\alpha\big(\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{I}_{0:k^*}} + (s+N)\gamma\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}}\big)}{\gamma(1-\gamma\alpha\,\mathrm{tr}(\mathbf{H}))} \cdot \bigg(\frac{k^*}{N^2} + \gamma^2\sum_{i>k^*}\lambda_i^2\bigg),\end{aligned}$$

where $k^* = \max\{k : \lambda_k \ge \frac{1}{\gamma N}\}$.

Proof Substituting the upper bound of $\mathbf{S}_{s:s+N}$ into (D.1), we can get

$$\begin{aligned}\mathrm{bias} \le \underbrace{\frac{\alpha\,\mathrm{tr}(\mathbf{B}_{s,s+N}+\mathbf{B}_{0,s})}{N^2(1-\gamma\alpha\,\mathrm{tr}(\mathbf{H}))}\sum_{k=0}^{N-1}\Big\langle \mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^N, (\mathbf{I}-\gamma\mathbf{H})^{2k}\mathbf{H}\Big\rangle}_{I_1} \\ + \underbrace{\frac{1}{\gamma N^2}\sum_{k=0}^{N-1}\Big\langle \mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^N, (\mathbf{I}-\gamma\mathbf{H})^{k+s}\mathbf{B}_0(\mathbf{I}-\gamma\mathbf{H})^{k+s}\Big\rangle}_{I_2}. \qquad \text{(D.6)}\end{aligned}$$

By (B.21), we can get the following bound on I_1,

$$I_1 \le \frac{\alpha\,\mathrm{tr}(\mathbf{B}_{s,s+N}+\mathbf{B}_{0,s})}{\gamma(1-\gamma\alpha\,\mathrm{tr}(\mathbf{H}))} \cdot \bigg(\frac{k^*}{N^2} + \gamma^2\sum_{i>k^*}\lambda_i^2\bigg). \qquad \text{(D.7)}$$

Then following the same procedure in (B.22), we have

$$\mathrm{tr}(\mathbf{B}_{s,s+N}+\mathbf{B}_{0,s}) \le 2\,\mathrm{tr}(\mathbf{B}_{0,s+N}) \le 4\big(\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{I}_{0:k^*}} + (s+N)\gamma\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}}\big)$$

where $k^* = \max\big\{k : \lambda_k \ge \frac{1}{\gamma N}\big\}$ (in fact k^* can be arbitrary chosen). Plugging this into (D.6) gives

$$I_1 \le \frac{4\alpha\big(\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{I}_{0:k^*}} + (s+N)\gamma\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}}\big)}{\gamma(1-\gamma\alpha\,\mathrm{tr}(\mathbf{H}))} \cdot \bigg(\frac{k^*}{N^2} + \gamma^2\sum_{i>k^*}\lambda_i^2\bigg).$$

Additionally, we have the following upper bound on I_2,

$$I_2 = \frac{1}{\gamma N^2}\sum_{k=0}^{N-1}\big\langle (\mathbf{I}-\gamma\mathbf{H})^{2(k+s)}\big(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^N\big), \mathbf{B}_0\big\rangle$$

$$\le \frac{1}{\gamma N^2}\sum_{k=0}^{N-1}\big\langle(\mathbf{I}-\gamma\mathbf{H})^{k+2s}-(\mathbf{I}-\gamma\mathbf{H})^{N+k+2s},\mathbf{B}_0\big\rangle.$$

Similar to the proof of Lemma B.11, let $\mathbf{v}_1,\mathbf{v}_2,\dots$ be the eigenvectors of $\mathbf{H}$ corresponding to its eigenvalues $\lambda_1,\lambda_2,\dots$ and $\omega_i=\mathbf{v}_i^\top(\mathbf{I}-\gamma\mathbf{H})^s(\mathbf{w}_0-\mathbf{w}^*)$, we have

$$\begin{aligned}
I_2 &\le \frac{1}{\gamma N^2}\sum_{k=0}^{N-1}\big\langle(\mathbf{I}-\gamma\mathbf{H})^{k}-(\mathbf{I}-\gamma\mathbf{H})^{N+k},(\mathbf{I}-\gamma\mathbf{H})^{2s}\mathbf{B}_0\big\rangle\\
&= \frac{1}{\gamma N^2}\sum_{k=0}^{N-1}\sum_i\big[(1-\gamma\lambda_i)^k-(1-\gamma\lambda_i)^{N+k}\big]\omega_i^2\\
&= \frac{1}{\gamma^2 N^2}\sum_i\frac{\omega_i^2}{\lambda_i}\big[1-(1-\gamma\lambda_i)^N\big]^2\\
&\le \frac{1}{\gamma^2 N^2}\sum_i\frac{\omega_i^2}{\lambda_i}\cdot\min\{1,\gamma^2N^2\lambda_i^2\}\\
&\le \frac{1}{\gamma^2 N^2}\cdot\sum_{i\le k^*}\frac{\omega_i^2}{\lambda_i}+\sum_{i>k^*}\lambda_i\omega_i^2\\
&= \frac{1}{\gamma^2 N^2}\cdot\big\|(\mathbf{I}-\gamma\mathbf{H})^s(\mathbf{w}_0-\mathbf{w}^*)\big\|_{\mathbf{H}_{0:k^*}^{-1}}^2+\big\|(\mathbf{I}-\gamma\mathbf{H})^s(\mathbf{w}_0-\mathbf{w}^*)\big\|_{\mathbf{H}_{k^*:\infty}}^2,
\end{aligned}\tag{D.8}$$

where $k^*=\max\{k:\lambda_k\ge\frac{1}{\gamma N}\}$. Combining (D.7) and (D.8) immediately completes the proof.

■

D.1.3 Proof of Theorem 5.1

Proof By Lemma D.2, it suffices to substitute into the upper bounds on the bias and variance errors. In particular, by Young's inequality we have

$$\mathbb{E}[L(\overline{\mathbf{w}}_N)]-L(\mathbf{w}^*)\le\Big(\sqrt{\mathrm{bias}}+\sqrt{\mathrm{variance}}\Big)^2\le 2\cdot\mathrm{bias}+2\cdot\mathrm{variance}.$$

Then we can directly substitute the bounds of variance and bias we proved in Lemmas D.3 and D.5. In particular, by Assumptions 2.2 we can directly get $R^2=\alpha\operatorname{tr}(\mathbf{H})$. Therefore, it holds that

$$\begin{aligned}
&\mathbb{E}[L(\overline{\mathbf{w}}_N)]-L(\mathbf{w}^*)\\
&\le 2\bigg[\frac{1}{\gamma^2N^2}\cdot\big\|(\mathbf{I}-\gamma\mathbf{H})^s(\mathbf{w}_0-\mathbf{w}^*)\big\|_{\mathbf{H}_{0:k^*}^{-1}}^2+\big\|(\mathbf{I}-\gamma\mathbf{H})^s(\mathbf{w}_0-\mathbf{w}^*)\big\|_{\mathbf{H}_{k^*:\infty}}^2\\
&\qquad+\frac{2\alpha\|\mathbf{w}_0-\mathbf{w}^*\|_2^2}{\gamma(1-\gamma\alpha\operatorname{tr}(\mathbf{H}))}\cdot\bigg(\frac{k^*}{N^2}+\gamma^2\sum_{i>k^*}\lambda_i^2\bigg)\\
&\qquad+\frac{\sigma^2}{1-\gamma\alpha\operatorname{tr}(\mathbf{H})}\cdot\bigg(\frac{k^*}{N}+\gamma\cdot\sum_{k^*<i\le k^\dagger}\lambda_i+\gamma^2(s+N)\cdot\sum_{i>k^\dagger}\lambda_i^2\bigg)\bigg]
\end{aligned}$$

$$= 2 \cdot \mathrm{EffectiveBias} + 2 \cdot \mathrm{EffectiveVar},$$

where

$$\begin{aligned}
\mathrm{EffectiveBias} &= \frac{1}{\gamma^2 N^2} \cdot \big\|(\mathbf{I} - \gamma\mathbf{H})^s(\mathbf{w}_0 - \mathbf{w}^*)\big\|^2_{\mathbf{H}^{-1}_{0:k^*}} + \big\|(\mathbf{I} - \gamma\mathbf{H})^s(\mathbf{w}_0 - \mathbf{w}^*)\big\|^2_{\mathbf{H}_{k^*:\infty}} \\
\mathrm{EffectiveVar} &= \frac{\sigma^2}{1 - \gamma\alpha\,\mathrm{tr}(\mathbf{H})} \cdot \left(\frac{k^*}{N} + \gamma \cdot \sum_{k^* < i \le k^\dagger} \lambda_i + \gamma^2 (s+N) \cdot \sum_{i > k^\dagger} \lambda_i^2\right) \\
&\quad + \frac{4\alpha\big(\|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{I}_{0:k^*}} + (s+N)\gamma\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}}\big)}{N\gamma(1 - \gamma\alpha\,\mathrm{tr}(\mathbf{H}))} \cdot \left(\frac{k^*}{N} + \gamma^2 N \sum_{i > k^*} \lambda_i^2\right).
\end{aligned}$$

■

D.2 Lower Bounds for Tail-Averaging

In this part we assume the noise is well-specified as in (2.1), and consider the SGD with tail-averaging

$$\overline{\mathbf{w}}_{s:s+N} = \frac{1}{N} \sum_{t=s}^{s+N} \mathbf{w}_t.$$

The following lemma is a variant of Lemma C.1, and lowers bound the excess risk.

Lemma D.6 (Variant of Lemma C.1) *Suppose the model noise ξ_t is well-specified, i.e., ξ_t and $\mathbf{x}_t$ are independent and $\mathbb{E}[\xi_t] = 0$. Then*

$$\begin{aligned}
\mathbb{E}[L(\overline{\mathbf{w}}_{s:s+N}) - L(\mathbf{w}^*)] \ge \frac{1}{2N^2} \cdot \sum_{t=0}^{N-1} \sum_{k=t}^{N-1} \Big\langle (\mathbf{I} - \gamma\mathbf{H})^{k-t}\mathbf{H}, \mathbf{B}_{s+t} \Big\rangle \\
+ \frac{1}{2N^2} \cdot \sum_{t=0}^{N-1} \sum_{k=t}^{N-1} \Big\langle (\mathbf{I} - \gamma\mathbf{H})^{k-t}\mathbf{H}, \mathbf{C}_{s+t} \Big\rangle.
\end{aligned}$$

We then present the lower bound for the variance error.

Lemma D.7 (Variant of Lemma C.3) *Suppose Assumptions 2.1 hold. Suppose the noise is well-specified (as in* (2.1)*). Suppose $N \ge 500$. Denote*

$$\mathrm{variance} = \frac{1}{2N^2} \cdot \sum_{t=0}^{N-1} \sum_{k=t}^{N-1} \Big\langle (\mathbf{I} - \gamma\mathbf{H})^{k-t}\mathbf{H}, \mathbf{C}_{s+t} \Big\rangle.$$

If the stepsize satisfies $\gamma < 1/\lambda_1$, then

$$\mathrm{variance} \ge \frac{\sigma^2_{noise}}{600} \left(\frac{k^*}{N} + \gamma \cdot \sum_{k^* < i \le k^\dagger} \lambda_i + (s+N)\gamma^2 \cdot \sum_{i > k^\dagger} \lambda_i^2\right),$$

where $k^* = \max\{k : \lambda_k \ge \frac{1}{N\gamma}\}$ and $k^\dagger = \max\{k : \lambda_k \ge \frac{1}{(s+N)\gamma}\}$.

Proof We can lower bound the variance error as follows

$$
\begin{aligned}
\text{variance} &= \frac{1}{2N^2}\sum_{t=0}^{N-1}\sum_{k=t}^{N-1}\big\langle(\mathbf{I}-\gamma\mathbf{H})^{k-t}\mathbf{H},\mathbf{C}_{s+t}\big\rangle \\
&= \frac{1}{2\gamma N^2}\sum_{t=0}^{N-1}\big\langle\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N-t},\mathbf{C}_{s+t}\big\rangle \\
&\ge \frac{\sigma^2_{\text{noise}}}{4N^2}\sum_{t=0}^{N-1}\big\langle\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N-t},\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{2(s+t)}\big\rangle \qquad \text{(use Lemma C.2)} \\
&= \frac{\sigma^2_{\text{noise}}}{4N^2}\sum_i\sum_{t=0}^{N-1}\left(1-(1-\gamma\lambda_i)^{N-t}\right)\left(1-(1-\gamma\lambda_i)^{2(s+t)}\right) \\
&\ge \frac{\sigma^2_{\text{noise}}}{4N^2}\sum_i\sum_{t=0}^{N-1}\left(1-(1-\gamma\lambda_i)^{N-t-1}\right)\left(1-(1-\gamma\lambda_i)^{s+t}\right),
\end{aligned}
$$

where $\{\lambda_i\}_{i\ge1}$ are the eigenvalues of $\mathbf{H}$ and are sorted in decreasing order. Define

$$
f(x) := \sum_{t=0}^{N-1}\left(1-(1-x)^{N-t-1}\right)\left(1-(1-x)^{s+t}\right), \qquad 0<x<1,
$$

then

$$
\text{variance} \ge \frac{\sigma^2_{\text{noise}}}{4N^2}\sum_i f(\gamma\lambda_i).
$$

We have the following lower bound for $f(x)$.

$$
\begin{aligned}
f(x) &= \sum_{t=0}^{N-1}\left(1-(1-x)^{N-t-1}\right)\left(1-(1-x)^{s+t}\right) \\
&\ge \sum_{t=\frac{N}{4}}^{\frac{3N}{4}-1}\left(1-(1-x)^{N-t-1}\right)\left(1-(1-x)^{s+t}\right) \\
&\ge \frac{N}{2}\left(1-(1-x)^{\frac{N}{4}}\right)\left(1-(1-x)^{s+\frac{N}{4}}\right)
\end{aligned}
$$

We then bound $f(x)$ by the range of x.

1. For $x>1/N$, we have that

$$
\begin{aligned}
f(x) &\ge \frac{N}{2}\left(1-(1-x)^{\frac{N}{4}}\right)\left(1-(1-x)^{\frac{N}{4}}\right) \\
&\ge \frac{N}{2}\left(1-\left(1-\frac{1}{N}\right)^{\frac{N}{4}}\right)\left(1-\left(1-\frac{1}{N}\right)^{\frac{N}{4}}\right) \\
&\ge \frac{N}{2}\left(1-\frac{1}{e^{1/4}}\right)\left(1-\frac{1}{e^{1/4}}\right) \ge \frac{N}{50}.
\end{aligned}
$$

2. For $1/N > x > 1/(s+N)$, we have that

$$\begin{aligned} f(x) &\ge \frac{N}{2}\left(1-(1-x)^{\frac{N}{4}}\right)\left(1-(1-x)^{\frac{s+N}{4}}\right)\\ &\ge \frac{N}{2}\left(1-(1-x)^{\frac{N}{4}}\right)\left(1-\left(1-\frac{1}{s+N}\right)^{\frac{s+N}{4}}\right)\\ &\ge \frac{N}{2}\left(1-\left(1-\frac{N}{8}x\right)\right)\left(1-\frac{1}{e^{1/4}}\right) \ge \frac{N^2 x}{100}. \end{aligned}$$

3. For $x < 1/(s+N) < 1/N$, we have that

$$\begin{aligned} f(x) &\ge \frac{N}{2}\left(1-(1-x)^{\frac{N}{4}}\right)\left(1-(1-x)^{s+\frac{N}{4}}\right)\\ &\ge \frac{N}{2}\left(1-\left(1-\frac{N}{8}x\right)\right)\left(1-\left(1-\frac{s+N/4}{2}x\right)\right)\\ &\ge \frac{(s+N)N^2}{128}x^2. \end{aligned}$$

In sum, we have that

$$f(x) \ge \begin{cases} \frac{N}{50}, & \frac{1}{N} \le x < 1,\\ \frac{N^2}{100}x, & \frac{1}{s+N} \le x < \frac{1}{N},\\ \frac{(s+N)N^2}{128}x^2, & 0 < x < \frac{1}{s+N}. \end{cases}$$

Set $k^* = \max\{k : \lambda_k \ge \frac{1}{N\gamma}\}$ and $k^\dagger = \max\{k : \lambda_k \ge \frac{1}{(s+N)\gamma}\}$, then

$$\begin{aligned} \text{variance} &\ge \frac{\sigma^2_{\text{noise}}}{4N^2}\sum_i f(\gamma\lambda_i)\\ &\ge \frac{\sigma^2_{\text{noise}}}{4N^2}\left(\frac{Nk^*}{50} + \frac{N^2}{100}\gamma\cdot\sum_{k^*<i\le k^\dagger}\lambda_i + \frac{(s+N)N^2}{128}\gamma^2\cdot\sum_{i>k^\dagger}\lambda_i^2\right)\\ &\ge \frac{\sigma^2_{\text{noise}}}{600}\left(\frac{k^*}{N} + \gamma\cdot\sum_{k^*<i\le k^\dagger}\lambda_i + (s+N)\gamma^2\cdot\sum_{i>k^\dagger}\lambda_i^2\right). \end{aligned}$$

This completes the proof.

■

Next we discuss the lower bound for the bias error. Similarly to (C.4) and using Lemma D.6, we have that

$$\begin{aligned} \text{bias} &\ge \frac{1}{2N^2}\sum_{t=0}^{N-1}\sum_{k=t}^{N-1}\left\langle(\mathbf{I}-\gamma\mathbf{H})^{k-t}\mathbf{H}, \mathbf{B}_{s+t}\right\rangle = \frac{1}{2\gamma N^2}\sum_{t=0}^{N-1}\left\langle\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N-t}, \mathbf{B}_{s+t}\right\rangle\\ &\ge \frac{1}{2\gamma N^2}\sum_{t=0}^{N/2}\left\langle\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N-t}, \mathbf{B}_{s+t}\right\rangle \end{aligned}$$

$$\geq \frac{1}{2\gamma N^2}\big\langle \mathbf{I} - (\mathbf{I}-\gamma\mathbf{H})^{N/2}, \sum_{t=0}^{N/2}\mathbf{B}_{s+t}\big\rangle. \tag{D.9}$$

Let $\mathbf{S}_{s:s+n} := \sum_{t=0}^{n-1}\mathbf{B}_{s+t} = \sum_{t=0}^{n-1}(\mathcal{I}-\gamma\mathcal{T})^t\circ\mathbf{B}_s$. We remain to build lower bound for $\mathbf{S}_{s:s+N/2+1}$. Comparing the definitions of $\mathbf{S}_{s:s+n}$ with $\mathbf{S}_n$, the only difference is that $\mathbf{B}_0$ is replaced by $\mathbf{B}_s$. Therefore we directly have the following lemma.

Lemma D.8 (Variant of Lemma C.4) *Suppose Assumptions 2.1 and 2.4 hold. If the stepsize satisfies $\gamma < 1/\lambda_1$, then for any $n \geq 2$, it holds that*

$$\mathbf{S}_{s:s+n} \succeq \frac{\beta}{4}\operatorname{tr}\Big(\Big(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{n/2}\Big)\mathbf{B}_s\Big)\cdot\Big(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{n/2}\Big) + \sum_{t=0}^{n-1}(\mathbf{I}-\gamma\mathbf{H})^t\cdot\mathbf{B}_s\cdot(\mathbf{I}-\gamma\mathbf{H})^t.$$

Lemma D.9 (Variant of Lemma C.5) *Suppose Assumptions 2.1 and 2.4 hold and $N \geq 2$. Denote*

$$\mathrm{bias} = \frac{1}{2N^2}\sum_{t=0}^{N-1}\sum_{k=t}^{N-1}\big\langle(\mathbf{I}-\gamma\mathbf{H})^{k-t}\mathbf{H}, \mathbf{B}_{s+t}\big\rangle,$$

then if the stepsize satisfies $\gamma < 1/\gamma_1$, it holds that

$$\begin{aligned}\mathrm{bias} \geq\ & \frac{1}{100\gamma^2N^2}\Big(\|(\mathbf{I}-\gamma\mathbf{H})^s(\mathbf{w}_0-\mathbf{w}^*)\|^2_{\mathbf{H}^{-1}_{0:k^*}} + \gamma^2N^2\,\|(\mathbf{I}-\gamma\mathbf{H})^s(\mathbf{w}_0-\mathbf{w}^*)\|^2_{\mathbf{H}_{k^*:\infty}}\Big)\\ &+ \frac{\beta\,\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{H}_{k^\dagger:\infty}}}{16000N}\cdot\Big(k^* + \gamma^2N^2\sum_{i>k^*}\lambda_i^2\Big),\end{aligned}$$

where $k^* = \max\{k : \lambda_k \geq \frac{1}{N\gamma}\}$ and $k^\dagger = \max\{k : \lambda_k \geq \frac{1}{(s+N)\gamma}\}$.

Proof According to (D.9) and Lemma D.8, we have that

$$\begin{aligned}\mathrm{bias} &\geq \frac{1}{2\gamma N^2}\big\langle\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N/2}, \mathbf{S}_{s:s+N/2+1}\big\rangle \geq \frac{1}{2\gamma N^2}\big\langle\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N/2}, \mathbf{S}_{s:s+N/2}\big\rangle\\ &\geq \underbrace{\frac{\beta}{8\gamma N^2}\operatorname{tr}\Big(\Big(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N/4}\Big)\mathbf{B}_s\Big)\cdot\big\langle\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N/2}, \mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N/4}\big\rangle}_{I_1}\\ &\quad+\underbrace{\frac{1}{2\gamma N^2}\Big\langle\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N/2}, \sum_{t=0}^{N/2-1}(\mathbf{I}-\gamma\mathbf{H})^t\cdot\mathbf{B}_s\cdot(\mathbf{I}-\gamma\mathbf{H})^t\Big\rangle}_{I_2}.\end{aligned}$$

Also noticing a lower bound for $\mathbf{B}_s$:

$$\mathbf{B}_s = (\mathcal{I}-\gamma\mathcal{T})^s\circ\mathbf{B}_0 \geq (\mathcal{I}-\gamma\widetilde{\mathcal{T}})^s\circ\mathbf{B}_0 = (\mathbf{I}-\gamma\mathbf{H})^s\cdot\mathbf{B}_0\cdot(\mathbf{I}-\gamma\mathbf{H})^s.$$

Then the first term is lower bounded by

$$I_1 \geq \frac{\beta}{8\gamma N^2}\operatorname{tr}\Big(\Big(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N/4}\Big)(\mathbf{I}-\gamma\mathbf{H})^{2s}\mathbf{B}_0\Big)\cdot\operatorname{tr}\Big(\Big(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N/4}\Big)^2\Big)$$

$$= \frac{\beta}{8\gamma N^2}\left(\sum_i \left(1-(1-\gamma\lambda_i)^{N/4}\right)(1-\gamma\lambda_i)^{2s}\omega_i^2\right)\cdot\left(\sum_i \left(1-(1-\gamma\lambda_i)^{N/4}\right)^2\right),$$

where $\omega_i = \mathbf{v}_i^\top(\mathbf{w}_0-\mathbf{w}^*)$ for $\mathbf{v}_1,\dots,\mathbf{v}_d$ being the eigenvectors of $\mathbf{H}$; and the second term is lower bounded by

$$\begin{aligned}
I_2 &= \frac{1}{2\gamma N^2}\left\langle \sum_{t=0}^{N/2-1}(\mathbf{I}-\gamma\mathbf{H})^{2t}\left(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N/2}\right),\mathbf{B}_s\right\rangle \\
&\ge \frac{1}{2\gamma N^2}\left\langle \sum_{t=0}^{N/2-1}(\mathbf{I}-2\gamma\mathbf{H})^{t}\left(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N/2}\right),\mathbf{B}_s\right\rangle \\
&\ge \frac{1}{4\gamma^2 N^2}\left\langle \left(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N/2}\right)^2\mathbf{H}^{-1},\mathbf{B}_s\right\rangle \\
&\ge \frac{1}{4\gamma^2 N^2}\left\langle \left(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N/4}\right)^2\mathbf{H}^{-1},(\mathbf{I}-\gamma\mathbf{H})^s\mathbf{B}_0(\mathbf{I}-\gamma\mathbf{H})^s\right\rangle \\
&= \frac{1}{4\gamma^2 N^2}\sum_i\left(1-(1-\gamma\lambda_i)^{N/4}\right)^2\lambda_i^{-1}\left((1-\gamma\lambda_i)^s\omega_i\right)^2.
\end{aligned}$$

To further lower bound the two terms, noticing the following inequalities:

$$1-(1-\gamma\lambda_i)^{\frac{N}{4}} \ge \begin{cases} 1-(1-\frac{1}{N})^{\frac{N}{4}} \ge 1-e^{-\frac{1}{4}} \ge \frac{1}{5}, & \lambda_i \ge \frac{1}{\gamma N}, \\ \frac{N}{4}\cdot\gamma\lambda_i - \frac{N(N-4)}{32}\cdot\gamma^2\lambda_i^2 \ge \frac{N}{5}\cdot\gamma\lambda_i, & \lambda_i < \frac{1}{\gamma N}, \end{cases}$$

and

$$(1-\gamma\lambda_i)^{2s} \ge \begin{cases} 0, & \lambda_i \ge \frac{1}{\gamma s}, \\ (1-\frac{1}{s})^{2s} \ge \frac{1}{16}, & \lambda_i < \frac{1}{\gamma s}, \end{cases}$$

where we use the fact that $(1-1/s)^{2s}$ is a strictly increase function of s for $s\ge 2$ so that $(1-1/s)^{2s}\ge 1/16$. Plugging these into the bounds for I_1 and I_2, and setting $k^* := \max\{k:\lambda_k \ge 1/(\gamma N)\}$ and $k^\dagger := \max\{k:\lambda_k\ge 1/(\gamma(s+N))\}$, we then obtain that

$$\begin{aligned}
I_1 &\ge \frac{\beta}{8\gamma N^2}\cdot\left(\frac{\gamma N}{80}\sum_{i>k^\dagger}\lambda_i\omega_i^2\right)\cdot\left(\frac{1}{25}\cdot k^* + \frac{\gamma^2N^2}{25}\cdot\sum_{i>k^*}\lambda_i^2\right) \\
&= \frac{\beta\left\|\mathbf{w}_0-\mathbf{w}^*\right\|_{\mathbf{H}_{k^\dagger:\infty}}^2}{16000N}\cdot\left(k^* + \gamma^2N^2\sum_{i>k^*}\lambda_i^2\right),
\end{aligned}$$

and that

$$\begin{aligned}
I_2 &\ge \frac{1}{4\gamma^2N^2}\left(\frac{1}{25}\cdot\sum_{i\le k^*}\lambda_i^{-1}\left((1-\gamma\lambda_i)^s\omega_i\right)^2 + \frac{\gamma^2N^2}{25}\cdot\sum_{i>K^*}\lambda_i\left((1-\gamma\lambda_i)^s\omega_i\right)^2\right) \\
&= \frac{1}{100\gamma^2N^2}\left(\left\|(\mathbf{I}-\gamma\mathbf{H})^s(\mathbf{w}_0-\mathbf{w}^*)\right\|_{\mathbf{H}_{0:k^*}^{-1}}^2 + \gamma^2N^2\left\|(\mathbf{I}-\gamma\mathbf{H})^s(\mathbf{w}_0-\mathbf{w}^*)\right\|_{\mathbf{H}_{k^*:\infty}}^2\right).
\end{aligned}$$

Summing up the two terms completes the proof. ■

D.2.1 Proof of Theorem 5.2

Proof Plugging the bounds of the bias error and variance error in Lemmas D.9 and D.7 into Lemma D.6 immediately completes the proof. ■