RISK BOUNDS OF ACCELERATED SGD FOR OVERPARAMETERIZED LINEAR REGRESSION

Xuheng Li[1], **Yihe Deng**[1], **Jingfeng Wu**[2], **Dongruo Zhou**[3], **Quanquan Gu**[1]
[1]Department of Computer Science, University of California, Los Angeles, CA 90095, USA
[2]Simons Institute, University of California, Berkeley, CA 94720, USA
[3]Department of Computer Science, Indiana University Bloomington, IN 47408, USA
xuheng.li@cs.cula.edu, yihedeng@cs.ucla.edu,
uuujf@berkeley.edu, dz13@iu.edu, qgu@cs.ucla.edu

ABSTRACT

Accelerated stochastic gradient descent (ASGD) is a workhorse in deep learning and often achieves better generalization performance than SGD. However, existing optimization theory can only explain the faster convergence of ASGD, but cannot explain its better generalization. In this paper, we study the generalization of ASGD for overparameterized linear regression, which is possibly the simplest setting of learning with overparameterization. We establish an instance-dependent excess risk bound for ASGD within each eigen-subspace of the data covariance matrix. Our analysis shows that (i) ASGD outperforms SGD in the subspace of small eigenvalues, exhibiting a faster rate of exponential decay for bias error, while in the subspace of large eigenvalues, its bias error decays slower than SGD; and (ii) the variance error of ASGD is always larger than that of SGD. Our result suggests that ASGD can outperform SGD when the difference between the initialization and the true weight vector is mostly confined to the subspace of small eigenvalues. Additionally, when our analysis is specialized to linear regression in the strongly convex setting, it yields a tighter bound for bias error than the best-known result.

1 INTRODUCTION

Momentum (Nesterov, 1983) is an important technique in optimization. In the context of convex and smooth optimization, Nesterov's momentum (accelerated gradient descent (AGD)) achieves the minimax optimal convergence rate (Nesterov, 2014) and provably accelerates the vanilla GD method. Recent work by Liu & Belkin (2018) shows that stochastic gradient descent (SGD) can also be accelerated by momentum in the overparameterized setting. However, the effect of momentum on the generalization performance is less studied. It has been empirically shown that ASGD does not always outperform SGD (Wang et al., 2023), but there has been little theoretical work justifying this observation. Notable exceptions are Jain et al. (2018) and Varre & Flammarion (2022), which provide excess risk bounds for accelerated SGD (ASGD) (a.k.a., SGD with momentum) for least squares problems in the strongly convex (Jain et al., 2018) and convex settings (Varre & Flammarion, 2022), respectively. However, both of their results are limited to the classical, finite-dimensional regime, and cannot be applied when the number of parameters exceeds the number of samples. On the other hand, a recent line of work completely characterizes the excess risk of SGD for least squares, even in the overparameterized regime (Dieuleveut & Bach, 2015; Défossez & Bach, 2015; Jain et al., 2017b; Berthier et al., 2020; Zou et al., 2021b; Wu et al., 2022). In particular, Zou et al. (2021b); Wu et al. (2022) provide finite-sample and dimension-free excess risk bounds for SGD that are sharp for each least squares instance. Given these results, it becomes imperative to thoroughly investigate whether the inclusion of momentum proves beneficial in terms of generalization, particularly in the context of least squares problems.

Contributions. In this paper, we tackle the question by considering ASGD for (overparameterized) linear regression problems and comparing its performance with SGD.

- Our main result provides an instance-dependent excess risk bound for ASGD that can be applied in the overparameterized regime. Similar to the bounds for SGD in Zou et al. (2021b); Wu et al. (2022), our bound for ASGD is independent of the ambient dimension and comprehensively depends on the spectrum of the data covariance matrix. When applied to the classical,

strongly-convex regime, our results recover the excess risk upper bounds in Jain et al. (2018), with significant improvements on the coefficient of the bias error.[1]

- Based on the excess risk bounds, we then compare the excess risk of ASGD and SGD. We find that the variance error of ASGD is always no smaller than that of SGD. Moreover, the bias error of ASGD is smaller than that of SGD along the small eigenvalue directions, but is larger than that of SGD along the large eigenvalue directions, with respect to the spectrum of the data covariance matrix. Thus momentum can help with generalization only if the main signals are aligned with small eigenvalue directions of the data covariance matrix and if the noise is small.
- From a technical perspective, we extend the analysis of the stationary covariance matrix in Jain et al. (2018) to the overparameterized setting, where we remove all dimension-dependent factors with a fine-grained analysis of the ASGD iterates. Our techniques might be of independent interest for analyzing ASGD in other settings.

Notation. In this paper, scalars are denoted by non-boldface letters. Vectors and matrices are denoted by lower-case and upper-case boldface letters, respectively. Denote linear operators on matrices by upper-case calligraphic letters. Denote the inner product of vectors by $\langle \mathbf{u}, \mathbf{v} \rangle$. For a vector $\mathbf{v}$, denote its j-th entry as $(\mathbf{v})_j$; For a matrix $\mathbf{M}$, denote its ij-entry as $(\mathbf{M})_{ij}$. For a PSD matrix $\mathbf{M}$, define $\|\mathbf{u}\|_{\mathbf{M}}^2 = \mathbf{u}^\top \mathbf{M} \mathbf{u}$. Denote the 2-norm of vector $\mathbf{v}$ as $\|\mathbf{v}\|_2 = \sqrt{\mathbf{v}^\top \mathbf{v}}$. Denote the inner product of matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{2d \times 2d}$ as $\langle \mathbf{A}, \mathbf{B} \rangle = \sum_{i,j=1}^{2d} (\mathbf{A})_{ij} (\mathbf{B})_{ij}$. The Kronecker product of matrices is denoted by $\otimes$. The operation of a linear matrix operator on a matrix is denoted by $\circ$.

2 Related Work

The generalization performances of SGD and ASGD applied to *underparameterized* linear regression have been studied in a line of works, based on the technique of bias-variance decomposition. It is shown that for SGD with iterate averaging from the beginning, bias error has a convergence rate of $\mathcal{O}(1/N^2)$ and variance has a convergence rate of $\mathcal{O}(d/N)$, where N is the number of calls of the stochastic oracle and d is the model dimension (Défossez & Bach, 2015; Dieuleveut et al., 2017; Jain et al., 2017a). If the eigenvalue of the data covariance matrix is bounded away from zero, then the convergence rate of the bias error can be further improved with additional exponential shrinkage by taking tail averaging of the iterates (Jain et al., 2017b).

For ASGD applied to linear regression, there are two cases: one with the assumption that the eigenvalue spectrum of the data covariance matrix is bounded away from zero (strongly convex) and the other without such assumption (general convex). For strongly convex linear regression, Jain et al. (2018) show an accelerated convergence rate for the bias error of ASGD with constant stepsize and tail averaging, compared to that of tail-averaged SGD in Jain et al. (2017b). We extend the use of linear operators and the techniques for bounding the operator spectrum in Jain et al. (2018).

Recently, the generalization of ASGD applied to general convex linear regression is studied by Varre & Flammarion (2022). Their result shows the acceleration of ASGD with time-varying parameters and weighted iterate averaging, especially for large N. The case of general convex linear regression is closer to the overparameterized setting where fast-decaying eigenspectrum is of special interest. However, their result is not applicable to the overparameterized linear regression because of the dimensionality dependence. Additionally, their result does not reveal the exponential bias decay of ASGD with constant stepsize.

The generalization performance of overparameterized linear regression has been studied by a line of works (Bartlett et al., 2020; Tsigler & Bartlett, 2020). For SGD applied to overparameterized linear regression, Zou et al. (2021b) replace the model dimensionality d with the effective dimension defined in terms of the eigenspectrum. This work manages to deal with any data covariance matrix, while prior works require certain assumptions (Dieuleveut & Bach, 2015). Wu et al. (2022) show a similar result for the last iterate of SGD with exponentially decaying stepsize.

3 Preliminaries

3.1 Linear Regression and ASGD

The goal of linear regression is to minimize the following risk:

$$L(\mathbf{w}) \coloneqq 1/2 \cdot \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \left[(y - \langle \mathbf{w}, \mathbf{x} \rangle)^2 \right],$$

[1] Our excess risk bound contains an extra term, which can be removed by a fine-grained analysis used by Jain et al. (2018) in the classical regime.

where $\mathbf{x}$ is an input feature vector belonging to a Hilbert space (denoted by $\mathcal{H}$, which could be either d-dimensional for a finite d, or countably infinite dimensional), $y \in \mathbb{R}$ is the response, $\mathbf{w} \in \mathcal{H}$ is the weight vector to be optimized, and $\mathcal{D}$ is an underlying unknown distribution of the data.
We consider the ASGD algorithm with tail averaging. In detail, in the t-th iteration, a sample $(\mathbf{x}_t, y_t) \sim \mathcal{D}$ is observed. Then the stochastic gradient is calculated by

$$\widehat{\nabla} L(\mathbf{w}) = -(y_t - \langle \mathbf{w}, \mathbf{x}_t \rangle)\mathbf{x}_t. \tag{3.1}$$

We follow the classical ASGD scheme (Nesterov, 2014), which maintains three sequences $\mathbf{w}_t$, $\mathbf{v}_t$ and $\mathbf{u}_t$. Let N be the number of samples observed, then for any $1 \le t \le N$, the update rules of $\mathbf{w}_t, \mathbf{v}_t, \mathbf{u}_t$ are as follows.

$$\mathbf{u}_{t-1} = \alpha \mathbf{w}_{t-1} + (1-\alpha)\mathbf{v}_{t-1}, \tag{3.2}$$

$$\mathbf{w}_t = \mathbf{u}_{t-1} - \delta \widehat{\nabla} L(\mathbf{u}_{t-1}), \tag{3.3}$$

$$\mathbf{v}_t = \beta \mathbf{u}_{t-1} + (1-\beta)\mathbf{v}_{t-1} - \gamma \widehat{\nabla} L(\mathbf{u}_{t-1}), \tag{3.4}$$

where $\alpha, \beta, \gamma, \delta > 0$ are hyperparameters. The $\mathbf{v}_t$ sequence is initialized at $\mathbf{w}_0 \in \mathcal{H}$. We remark that ASGD reduces to stochastic heavy ball (SHB, Polyak (1964)) when $\delta = 0$, so our results can be directly applied to SHB by setting $\delta = 0$ (see Appendix C for details). We also remark that ASGD reduces to SGD when $\delta = \gamma$.
In this work, following Jain et al. (2018) and Zou et al. (2021b), we consider ASGD with tail averaging. The tail-averaged final output is $\overline{\mathbf{w}}_{s,s+N} := N^{-1} \sum_{t=s}^{s+N-1} \mathbf{w}_t$. With certain assumptions, $L(\mathbf{w})$ admits a unique global optimum denoted by $\mathbf{w}^* := \operatorname{argmin}_{\mathbf{w}} L(\mathbf{w})$. We focus on the overparameterized setting, where $d \gg N$ (or possibly countably infinite).
Define the centered ASGD iterate as $\boldsymbol{\eta}_t := \begin{bmatrix} \mathbf{w}_t - \mathbf{w}^* \\ \mathbf{u}_t - \mathbf{w}^* \end{bmatrix}$. Denote the noise in each sample as $\epsilon_t := y_t - \langle \mathbf{w}^*, \mathbf{x}_t \rangle$. By (3.1), the stochastic gradient at $\mathbf{u}_{t-1}$ can be expressed as

$$\widehat{\nabla} L(\mathbf{u}_{t-1}) = -(\epsilon_t + \langle \mathbf{w}^*, \mathbf{x}_t \rangle - \langle \mathbf{u}_{t-1}, \mathbf{x}_t \rangle)\mathbf{x}_t = \mathbf{x}_t \mathbf{x}_t^\top (\mathbf{u}_{t-1} - \mathbf{w}^*) - \epsilon_t \mathbf{x}_t. \tag{3.5}$$

By substituting (3.5) into (3.3) and (3.4) and eliminating $\mathbf{v}_t$ using (3.2), we have

$$\boldsymbol{\eta}_t = \widehat{\mathbf{A}}_t \boldsymbol{\eta}_{t-1} + \boldsymbol{\zeta}_t, \quad \text{where} \quad \widehat{\mathbf{A}}_t := \begin{bmatrix} \mathbf{0} & \mathbf{I} - \delta \mathbf{x}_t \mathbf{x}_t^\top \\ -c\mathbf{I} & (1+c)\mathbf{I} - q \mathbf{x}_t \mathbf{x}_t^\top \end{bmatrix}, \quad \boldsymbol{\zeta}_t := \begin{bmatrix} \delta \cdot \epsilon_t \mathbf{x}_t \\ q \cdot \epsilon_t \mathbf{x}_t \end{bmatrix},$$

and $c := \alpha(1-\beta), q := \alpha\delta + (1-\alpha)\gamma$. Denote the expectation of $\widehat{\mathbf{A}}_t$ as

$$\mathbf{A} := \mathbb{E}[\widehat{\mathbf{A}}_t] = \begin{bmatrix} \mathbf{0} & \mathbf{I} - \delta \mathbf{H} \\ -c\mathbf{I} & (1+c)\mathbf{I} - q\mathbf{H} \end{bmatrix},$$

where $\mathbf{H} = \mathbb{E}_{\mathbf{x} \sim \mathcal{D}|_{\mathbf{x}}}[\mathbf{x}\mathbf{x}^\top]$ is the second-order moment matrix of the distribution $\mathcal{D}$, which is also the Hessian of $L(\mathbf{w})$. Let the eigen-decomposition of the Hessian be $\mathbf{H} = \sum_{i=1}^d \lambda_i \mathbf{v}_i \mathbf{v}_i^\top$, where $\{\lambda_i\}_{i=1}^d$ are the eigenvalues of $\mathbf{H}$ sorted in descending order with $\mathbf{v}_i$'s being the corresponding eigenvectors. Similar to Jain et al. (2018), we assume that $\mathbf{H}$ is diagonal, then $\mathbf{A}$ is block diagonal with each block being $\mathbf{A}_i := \begin{bmatrix} 0 & 1 - \delta\lambda_i \\ -c & 1 + c - q\lambda_i \end{bmatrix}$. In this work, we are particularly interested in analyzing the eigenvalues of $\mathbf{A}_i$, since the spectral norm of $\mathbf{A}_i$ determines the decay rate of the bias error in the subspace of λ_i.

3.2 Assumptions

We then introduce assumptions required in our analysis, following those of Zou et al. (2021b); Wu et al. (2022). Our first assumption regularizes the moments of the data distribution.

Assumption 3.1 (Regularity conditions)**.** The second moment $\mathbf{H}$ exists, and $\operatorname{tr}(\mathbf{H})$ is finite. $\mathbf{H}$ is strictly positive definite, i.e., $\mathbf{H} \succ \mathbf{0}$. Thus, $L(\mathbf{w})$ admits a unique global optimum $\mathbf{w}^*$. The second-order moment of labels $\mathbb{E}[y^2]$ is also finite. Let $\mathcal{M}$ denote the fourth moment of $\mathbf{x}$:

$$\mathcal{M} := \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{x} \otimes \mathbf{x} \otimes \mathbf{x} \otimes \mathbf{x}].$$

Then $\mathcal{M}$ exists and is finite.

Our second assumption is a proposition of the fourth moment of $\mathbf{x}$, viewed as a linear operator $\mathcal{M}$ on PSD matrices.

Assumption 3.2 (Fourth moment condition)**.** Assume there exists a positive constant $\psi > 0$, such that for any PSD matrix $\mathbf{A}$, it holds that

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

A special case of Assumption 3.2 is when $\mathcal{D}$ is a Gaussian distribution. For that case, we have $\psi = 3$. We remark that although Assumption 3.2 does not cover some special cases, e.g., the one-hot distribution discussed in Zou et al. (2021a), similar results can still be obtained by applying our techniques with minor modifications (see Appendix J for details).
The following assumption characterizes the noise of the stochastic gradient.

Assumption 3.3 (Noise condition)**.** Assume that

$$\mathbf{\Sigma} \coloneqq \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\widehat{\nabla}L(\mathbf{w}^*)\otimes\widehat{\nabla}L(\mathbf{w}^*)] = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-\langle\mathbf{w}^*,\mathbf{x}\rangle)^2\mathbf{x}\mathbf{x}^\top],$$

and $\sigma^2 \coloneqq \|\mathbf{H}^{-\frac{1}{2}}\mathbf{\Sigma}\mathbf{H}^{-\frac{1}{2}}\|_2$ exist and are finite. Here, $\mathbf{\Sigma}$ is the covariance matrix of the gradient noise at $\mathbf{w}^*$. For *well-specified models* where $y_t - \langle\mathbf{w}^*,\mathbf{x}_t\rangle \sim \mathcal{N}(0,\sigma^2_{\text{noise}})$, we have $\mathbf{\Sigma} = \sigma^2_{\text{noise}}\mathbf{H}$ and thus $\sigma^2 = \sigma^2_{\text{noise}}$.

4 Main Results

We now provide an excess risk upper bound for ASGD.

4.1 Risk Bound of ASGD in the High-Dimensional Setting

Before we present the results, we first introduce three quantities which are cutoffs of the spectrum of $\mathbf{H}$. The eigenvalues of $\mathbf{A}_i$ can be either complex or real, which depends on the range of λ_i. Define

$$\begin{aligned} k^\ddagger &\coloneqq \max\{i : \lambda_i \geq (\sqrt{q-c\delta}+\sqrt{c(q-\delta)})^2/q^2\}, \\ k^\dagger &\coloneqq \max\{i : \lambda_i > (\sqrt{q-c\delta}-\sqrt{c(q-\delta)})^2/q^2\}. \end{aligned} \tag{4.1}$$

It is easy to see that $k^\ddagger \leq k^\dagger$. For any $i \leq k^\ddagger$ and any $i > k^\dagger$, $\mathbf{A}_i$ has real eigenvalues $x_1 \leq x_2$, and for i between $k^\ddagger$ and $k^\dagger$, $\mathbf{A}_i$ has complex eigenvalues x_1, x_2 with the same magnitude. We also define $\widehat{k}$ as

$$\widehat{k} \coloneqq \max\left\{i : \lambda_i \geq (1-c)/\delta\right\}.$$

Parameter choice. We select hyperparameters of ASGD as follows: We first pick a non-negative integer $\widetilde{\kappa}$. We then select parameters $\delta, \gamma, \beta, \alpha$ as follows, based on $\widetilde{\kappa}$:

$$\delta \leq \frac{1}{2\psi\,\mathrm{tr}(\mathbf{H})}, \quad \gamma \in \left[\delta,\ \frac{1}{2\psi\sum_{i>\widetilde{\kappa}}\lambda_i}\right], \quad \beta = \frac{\delta}{\psi\widetilde{\kappa}\gamma}, \quad \alpha = \frac{1}{1+\beta}. \tag{4.2}$$

We can show that with our choice of paramters, we have $k^\ddagger \leq \widehat{k} \leq k^\dagger$ (see Appendix E.1 for details). For convenience, we introduce the following notations for submatrices of $\mathbf{H}$: for any non-negative integers $k_1 \leq k_2$, denote

$$\mathbf{H}_{k_1:k_2} \coloneqq \sum_{i=k_1+1}^{k_2}\lambda_i\mathbf{v}_i\mathbf{v}_i^\top, \quad \mathbf{H}_{k_1:\infty} \coloneqq \sum_{i=k_1+1}^{d}\lambda_i\mathbf{v}_i\mathbf{v}_i^\top.$$

Now we present the main result, which gives a finite excess risk bound for ASGD under the specific parameter choice (4.2).

Theorem 4.1. Under Assumptions 3.1, 3.2 and 3.3, with the parameter choice in (4.2), if $N(1-c) \geq 2$, the excess risk of tail-averaged iterate from ASGD satisfies:

$$\mathbb{E}[L(\overline{\mathbf{w}}_{s,s+N})] - L(\mathbf{w}^*) \leq 2\cdot\text{EffectiveVar} + 2\cdot\text{EffectiveBias}. \tag{4.3}$$

where the effective variance is bounded by

$$\begin{aligned} \text{EffectiveVar} \leq \sigma^2 r\left[\frac{27k^*}{2N} + 18(s+N)\gamma^2\sum_{i>k^*}\lambda_i^2\right] + \frac{\psi r}{N}\left[\frac{9k^*}{N} + 36N\gamma^2\sum_{i>k^*}\lambda_i^2\right]\cdot\left[\frac{14}{\delta}\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{I}_{0:\widehat{k}}}\right. \\ \left. + \frac{10}{1-c}\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{H}_{\widehat{k}:k^\dagger}} + \frac{2}{\gamma+\delta}\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{I}_{k^\dagger:k^*}} + 4(s+N)\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}}\right], \end{aligned}$$

and the effective bias is bounded by

$$\begin{aligned}\text{EffectiveBias} \leq\ & \frac{8(c\delta/q)^{2s}}{N^2\delta^2}\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k^\ddagger}} + \frac{4s^2}{N^2}c^s\|(\mathbf{I}-\delta\mathbf{H})^{s/2}(\mathbf{w}_0-\mathbf{w}^*)\|^2_{\mathbf{H}_{k^\ddagger:k^\dagger}} \\ &+ \frac{16c^s}{N^2\delta^2}\|(\mathbf{I}-\delta\mathbf{H})^{s/2}(\mathbf{w}_0-\mathbf{w}^*)\|^2_{\mathbf{H}^{-1}_{k^\ddagger:\widehat{k}}} + \frac{100c^s}{N^2(1-c)^2}\|(\mathbf{I}-\delta\mathbf{H})^{s/2}(\mathbf{w}_0-\mathbf{w}^*)\|^2_{\mathbf{H}_{\widehat{k}:k^\dagger}} \\ &+ \frac{18}{N^2(\gamma+\delta)^2}\Big\|\Big(\mathbf{I}-\frac{\gamma+\delta}{2}\mathbf{H}\Big)^s(\mathbf{w}_0-\mathbf{w}^*)\Big\|^2_{\mathbf{H}^{-1}_{k^\dagger:k^*}} + 18\Big\|\Big(\mathbf{I}-\frac{\gamma+\delta}{2}\mathbf{H}\Big)^s(\mathbf{w}_0-\mathbf{w}^*)\Big\|^2_{\mathbf{H}_{k^*:\infty}},\end{aligned}$$

with $k^* = \max\{k : \lambda_k \geq 1/((\gamma+\delta)N)\}$, and

$$r \coloneqq \frac{1}{1-\psi l}, \quad l \coloneqq \frac{\delta\,\mathrm{tr}(\mathbf{H})}{2} + \frac{1}{2\psi} + \frac{\gamma}{4}\sum_{i>\widetilde{\kappa}}\lambda_i.$$

Theorem 4.1 establishes the excess risk bound of ASGD under the overparameterized setting. To our knowledge, this is the first instance-dependent bound of ASGD within each eigen-subspace of $\mathbf{H}$. Our excess bound includes both the variance term, which depends on the randomness coming from the data distribution $\mathcal{D}$, and the bias term, which includes "accelerated convergence" terms brought by the ASGD.

Remark 4.2. The cutoff index k^* is referred to as the *effective dimension*, which can be much smaller than the model dimensionality d, especially when the eigenvalues decay fast. We want to emphasize that similar effective dimension has also appeared in the previous work which analyzes the convergence of SGD under the overparameterized model setting (Zou et al., 2021b; Wu et al., 2022). Nevertheless, the effective dimension of SGD is $k^*_{\mathrm{SGD}} \coloneqq \max\{k : \lambda_k \geq 1/(\delta N)\}$, which is smaller than that in ASGD. In Section 5, we will provide a comparison of the risk bounds between SGD and ASGD.

Remark 4.3. It is worth noting that under the parameter selection (4.2), one can verify that $\psi l < 1$. Such a condition guarantees that $r = 1/(1-\psi l)$ is finite, which further guarantees that our derived risk bound for effective variance is valid.

4.2 Implication in the Classical Setting

In this subsection, we show that Theorem 4.1 implies the excess risk bound in the strongly convex setting and can recover a similar result as Jain et al. (2018). The hyperparameters of ASGD are chosen to be

$$\delta = \frac{1}{2\psi\,\mathrm{tr}(\mathbf{H})}, \quad \gamma = \sqrt{\frac{2\delta}{\psi\mu d}}, \quad \beta = \sqrt{\frac{\mu\delta}{2\psi d}}, \quad \alpha = \frac{1}{1+\beta}, \tag{4.4}$$

where $\mu \coloneqq \lambda_d$ is the smallest eigenvalue of $\mathbf{H}$. We remark that the parameter choice in (4.4) is different from the choice under the overparameterized setting given in (4.2) because $\widetilde{\kappa}$ is chosen as the model dimension d, and the upper bound of γ in (4.2), which is $1/(2\psi\sum_{i>\widetilde{\kappa}}\lambda_i)$, becomes vacuous. Instead, we require $\gamma = 2\beta/\mu$ to guarantee that no eigenvalue falls in the region of small eigenvalues such that $\mathbf{A}_i$ has real eigenvalues (i.e., when $i > k^\dagger$, see Section I for detailed proof). The following corollary provides the excess risk bound in the strongly convex setting:

Corollary 4.4. Under Assumptions 3.1, 3.2 and 3.3, and with the parameter choice in (4.4), the excess risk of tail-averaged iterate from ASGD in the classical regime satisfies:

$$\begin{aligned}\mathbb{E}[L(\overline{\mathbf{w}}_{s:s+N})] - L(\mathbf{w}^*) \leq\ & \underbrace{\frac{100}{N^2\beta^2}\exp\Big(-\frac{\beta s}{2}\Big)[L(\mathbf{w}_0)-L(\mathbf{w}^*)]}_{\text{Effective Bias}} \\ &+ \underbrace{\frac{1008\psi d}{N^2\beta}[L(\mathbf{w}_0)-L(\mathbf{w}^*)] + \frac{36\sigma^2 d}{N} + \frac{128\sigma^2 d}{N^2\beta}}_{\text{Effective Variance}}.\end{aligned}$$

Denote $\kappa \coloneqq \mathrm{tr}(\mathbf{H})/\mu$, then $\beta = \Theta(1/\sqrt{\kappa\widetilde{\kappa}})$. Assuming that $L(\mathbf{w}_0) - L(\mathbf{w}^*) = \mathcal{O}(\sigma^2)$, then the bound given in Corollary 4.4 fully recovers the excess risk upper bound given in Theorem 1 of Jain et al. (2018) in terms of exponential decay rate, leading-order variance and lower-order variance. Moreover, the coefficient of effective bias is $\mathcal{O}(\kappa\widetilde{\kappa}/N^2)$, which significantly improves upon

$\mathcal{O}(\kappa^{13/4}\widetilde{\kappa}^{9/4}d/N^2)$ given in Jain et al. (2018). It is worth noting that Liu & Belkin (2018) proved $\mathcal{O}(1)$ coefficient for effective bias of ASGD. Our result can also recover the constant coefficient when $N(1-c) \geq 2$, because $1-c = 2\alpha\beta \leq 2\beta$ and $1/(N^2\beta^2) \leq 1$. The difference in this coefficient between the bound in Liu & Belkin (2018) and ours is mainly due to slightly different treatments of terms in the form of $N^{-1}\sum_{i=0}^{N-1}(1-\beta)^i$, which is not essential.

5 Comparison between ASGD and SGD

In this section, we first introduce the SGD update, which is given by

$$\mathbf{w}_t^{\mathrm{SGD}} = \mathbf{w}_{t-1}^{\mathrm{SGD}} - \delta\widehat{\nabla}L(\mathbf{w}_{t-1}^{\mathrm{SGD}}),$$

where δ satisfies the requirement in (4.2). Analogous to ASGD, tail-averaged SGD is defined as $\overline{\mathbf{w}}_{s:s+N}^{\mathrm{SGD}} \coloneqq N^{-1}\sum_{t=s}^{s+N-1}\mathbf{w}_t^{\mathrm{SGD}}$. The excess risk of tail-averaged SGD is then $\mathbb{E}[L(\overline{\mathbf{w}}_{s:s+N}^{\mathrm{SGD}})] - L(\mathbf{w}^*)$. We then present the following theorem, which shows the existence of linear regression instances where ASGD outperforms SGD (the proof is given in Appendix D.2):

Theorem 5.1 (Informal)**.** There exists a class of linear regression instances and corresponding choice of parameter such that the excess risk bound of tail-averaged ASGD satisfies

$$\mathbb{E}[L(\overline{\mathbf{w}}_{s:s+N})] - L(\mathbf{w}^*) = \mathcal{O}(\sigma^2(N^{-1/2} + N^{-2}\cdot 0.9873^s)),$$

and the excess risk bound of tail-averaged SGD satisfies

$$\mathbb{E}[L(\overline{\mathbf{w}}_{s:s+N}^{\mathrm{SGD}})] - L(\mathbf{w}^*) = \Omega(\sigma^2(N^{-1/2} + N^{-2}\cdot 0.996^s)).$$

Theorem 5.1 is inspired by the following comparison of the effective variance and bias of SGD and ASGD with the assumption that $s = \mathcal{O}(N)$. This is a technical assumption that helps to simplify excess risk bounds, and the comparison can be extended to the case of $s = \Omega(N)$. Under the same set of assumptions as Theorem 4.1, Zou et al. (2021b) prove that, with a bias-variance decomposition similar to (4.3), effective variance and effective bias of SGD satisfy:

$$\text{EffectiveVar} \leq \sigma^2 r_{\mathrm{SGD}}\cdot\left[\frac{k_{\mathrm{SGD}}^*}{N} + (s+N)\delta^2\sum_{i>k_{\mathrm{SGD}}^*}\lambda_i^2\right]$$
$$+ \frac{4\psi r_{\mathrm{SGD}}}{N}\cdot\left[\frac{1}{\delta}\|\mathbf{w}_0-\mathbf{w}^*\|_{\mathbf{I}_{0:k_{\mathrm{SGD}}^*}}^2 + (s+N)\|\mathbf{w}_0-\mathbf{w}^*\|_{\mathbf{H}_{k_{\mathrm{SGD}}^*:\infty}}^2\right]\cdot\left[\frac{k_{\mathrm{SGD}}^*}{N} + N\delta^2\sum_{i>k_{\mathrm{SGD}}^*}\lambda_i^2\right],$$

$$\text{EffectiveBias} \leq \frac{1}{\delta^2N^2}\|(\mathbf{I}-\delta\mathbf{H})^s(\mathbf{w}_0-\mathbf{w}^*)\|_{\mathbf{H}_{0:k_{\mathrm{SGD}}^*}^{-1}}^2 + \|(\mathbf{I}-\delta\mathbf{H})^s(\mathbf{w}_0-\mathbf{w}^*)\|_{\mathbf{H}_{k_{\mathrm{SGD}}^*:\infty}}^2,$$

where $r_{\mathrm{SGD}} = (1-\psi\delta\,\mathrm{tr}(\mathbf{H}))^{-1}$ and $k_{\mathrm{SGD}}^* = \max\{i : \lambda_i \geq 1/(\delta N)\}$.

Comparison of effective variance. Assuming that the initial variance $\mathbf{w}_0 - \mathbf{w}^*$ is bounded, the effective variance of ASGD is dominated by

$$\sigma^2 r\left[\frac{24k^*}{N} + 18(s+N)\gamma^2\sum_{i>k^*}\lambda_i^2\right],$$

and effective variance of SGD is dominated by

$$\sigma^2 r_{\mathrm{SGD}}\left[\frac{k_{\mathrm{SGD}}^*}{N} + (s+N)\delta^2\sum_{i>k_{\mathrm{SGD}}^*}\lambda_i^2\right].$$

Thus, ignoring σ^2, r and r_{SGD} and constants, effective variance of ASGD in the subspace of λ_i is $\mathcal{O}(\min\{1/N, N\gamma^2\lambda_i^2\})$, compared to $\mathcal{O}(\min\{1/N, N\delta^2\lambda_i^2\})$ for SGD. With $\gamma \geq \delta$ according to the choice of parameters in (4.2), we conclude that the excess variance of ASGD in every subspace is larger than that of SGD.

The following corollary characterizes the effective variance of ASGD when the eigenvalue spectrum decays with a polynomial or exponential rate. These examples have been studied for SGD in Zou et al. (2021b) and Wu et al. (2022).

Corollary 5.2. Under the same assumptions as Theorem 4.1, suppose that $\|\mathbf{w}_0 - \mathbf{w}^*\|_2$ is bounded.

1. If the spectrum is $\lambda_i = i^{-(1+r)}$ for some $r > 0$, then the effective variance is $\mathcal{O}((\widetilde{\kappa}/N)^{r/(1+r)})$.

2. If the spectrum is $\lambda_i = e^{-i}$, then the effective variance is $\mathcal{O}((\widetilde{\kappa} + \log N)/N)$.

Remark 5.3. For SGD, the effective variance is $\mathcal{O}((1/N)^{r/(1+r)})$ if the eigenvalue spectrum is $\lambda_i = i^{-(1+r)}$, and $\mathcal{O}(\log N/N)$ if the eigenvalue spectrum is $\lambda_i = e^{-i}$ (Zou et al., 2021b). Therefore, the effective variance of ASGD is larger than that of SGD under both eigenvalue spectra.

Comparison of effective bias. Effective bias of both SGD and ASGD decay exponentially in s within each subspace. The decay rate of SGD is $(1-\delta\lambda_i)^s$ in the subspace of λ_i. For ASGD,

1. When $i \le k^{\ddagger}$, the decay rate in the subspace of λ_i is $(c\delta/q)^s$. By definition of $k^{\ddagger}$, we have $1-\delta\lambda_i \le c\delta/q$ (see Appendix E.1 for detailed proof).

2. When $k^{\ddagger} < i \le k^{\dagger}$, the decay rate in the subspace of λ_i is $[c(1-\delta\lambda_i)]^{s/2}$. According to the definition of $\widehat{k}$, when $k^{\ddagger} < i \le \widehat{k}$, we have $1-\delta\lambda_i \le \sqrt{c(1-\delta\lambda_i)}$; When $\widehat{k} < i \le k^{\dagger}$, we have $1-\delta\lambda_i \ge \sqrt{c(1-\delta\lambda_i)}$.

3. When $i > k^{\dagger}$, the decay rate in the subspace of λ_i is $(1-(\gamma+\delta)\lambda_i/2)^s$. By the choice of parameters (4.2), we have $\gamma \ge \delta$, so $1-(\gamma+\delta)\lambda_i/2 \le 1-\delta\lambda_i$.

Combining the three cases above, we conclude that the effective bias of ASGD decays faster than that of SGD in eigen-subspaces of λ_i where $i > \widehat{k}$, while it decays slower than SGD in subspaces of λ_i where $i \le \widehat{k}$. This phenomenon is illustrated in Figure 1. Therefore, ASGD can perform better than SGD if $\mathbf{w}_0 - \mathbf{w}^*$ is mostly refined to the eigen-subspaces of λ_i where $i > \widehat{k}$.

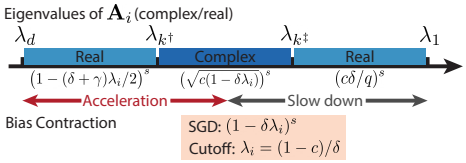


Figure 1: Illustration of the eigenspectrum.

We remark that this result is consistent with the acceleration of bias decay presented in Jain et al. (2018). Without instance-specific analysis, the exponential decay rate of bias is determined by the decay rate in subspace of the smallest eigenvalue. As the effective bias of ASGD decays faster than that of SGD in the eigen-subspace of small eigenvalues, the worst-case decay rate of the bias error of ASGD enjoys acceleration compared to SGD.

6 Experiments

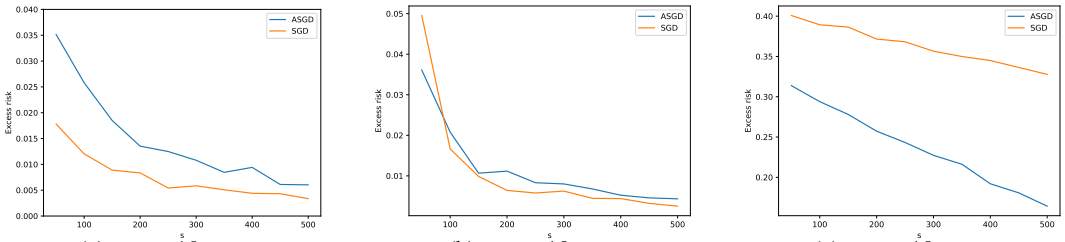
In this section, we empirically verify that ASGD can outperform SGD when $\mathbf{w}_0 - \mathbf{w}^*$ is mainly confined to the eigen-subspace of small eigenvalues.

Data model. Our experiments are based on the setting of overparameterized linear regression, where the model dimenstion is $d = 2000$. The data covariance matrix $\mathbf{H}$ is diagonal with eigenvalues $\lambda_i = i^{-2}$. The input $\mathbf{x}_t$ follows Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{H})$, so Assumption 3.2 holds with $\psi = 3$. The ground truth weight vector is $\mathbf{w}^* = \mathbf{0}$, and the label y_t follows the distribution $\mathcal{N}(0, \sigma^2)$ where $\sigma^2 = 0.01$.

Hyperparameters of ASGD and SGD. We select parameters of ASGD so that it satisfies the requirements in (4.2). We first let $\widetilde{\kappa} = 5$. According to (4.2), δ satisfies $\delta \le 1/\pi^2$, so we pick $\delta = 0.1$, which is also the stepsize of SGD. We then let $\alpha = 0.9875$, so that $(1-c)/\delta = 2(1-\alpha)/\delta = 0.25 = \lambda_2$, which implies that $\widehat{k} = 2$. Finally, we select $\beta = (1-\alpha)/\alpha$ and $\gamma = \delta/(\psi\widetilde{\kappa}\beta)$. We can verify that the parameters satisfy all requirements in (4.2).

We fix the length of tail averaging as $N = 500$, and conduct experiments on different s where $s = 50, 100, 150, \ldots, 500$. In each experiment, we measure $\overline{\mathbf{w}}_{s:s+N}^{\top}\mathbf{H}\overline{\mathbf{w}}_{s:s+N}$. For each s, we run the experiment 10 times and take the average of the test results.

We examine three different initializations: (a) $\mathbf{w}_0 = 10 \cdot \mathbf{e}_1$, representing the case where $\mathbf{w}_0 - \mathbf{w}^*$ is mainly refined to the subspace of large eigenvalues, (b) $\mathbf{w}_0 = 10 \cdot \mathbf{e}_2$, representing the case where $\mathbf{w}_0 - \mathbf{w}^*$ is mainly refined to the subspace of $\lambda_{\widehat{k}}$, and (c) $\mathbf{w}_0 = 10 \cdot \mathbf{e}_{20}$, representing the case where $\mathbf{w}_0 - \mathbf{w}^*$ is mainly refined to the subspace of small eigenvalues. Experiment results are shown in Figure 2. We observe that ASGD indeed outperforms SGD in the scenario where $\mathbf{w}_0 - \mathbf{w}^*$ is mostly refined to the subspace of small eigenvalues, and performs worse than SGD when $\mathbf{w}_0 - \mathbf{w}^*$ is refined to the subspace of large eigenvalues. Additionally, the excess risks of SGD and ASGD are similar when $\mathbf{w}_0 - \mathbf{w}^*$ aligns with the subspace corresponding to $\lambda_{\widehat{k}}$, which is also aligns with the implication of Theorem 4.1.



(a) $\mathbf{w}_0 = \sqrt{10} \cdot \mathbf{e}_1$. (b) $\mathbf{w}_0 = \sqrt{10} \cdot \mathbf{e}_2$. (c) $\mathbf{w}_0 = \sqrt{10} \cdot \mathbf{e}_{20}$.

Figure 2: Comparison of excess risk of ASGD and SGD. The noise scale is $\sigma^2 = 0.01$. We run each experiment 10 times and take the average of the excess risk in the 10 trials.

7 Proof Sketch

In this section, we present the high-level ideas in our proof. We mainly introduce two main ideas of the proof, including (i) bias-variance decomposition, and (ii) analysis of excess risk bounds within each eigen-subspace, based on the eigenvalues of $\mathbf{A}_i$.

Define the tail averaged centered ASGD iterate as $\overline{\boldsymbol{\eta}}_{s,s+N} \coloneqq N^{-1} \sum_{t=s}^{s+N-1} \boldsymbol{\eta}_t$. The excess risk is then

$$\mathbb{E}[L(\overline{\mathbf{w}}_{s,s+N})] - L(\mathbf{w}^*) = \frac{1}{2} \left\langle \begin{bmatrix} \mathbf{H} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \mathbb{E}[\overline{\boldsymbol{\eta}}_{s,s+N} \otimes \overline{\boldsymbol{\eta}}_{s,s+N}] \right\rangle.$$

We also define the linear operators $\mathcal{B} \coloneqq \mathbb{E}[\widehat{\mathbf{A}}_t \otimes \widehat{\mathbf{A}}_t]$ and $\widetilde{\mathcal{B}} \coloneqq \mathbf{A} \otimes \mathbf{A}$, which are both PSD operators. Additionally, the difference $\mathcal{B} - \widetilde{\mathcal{B}}$ is also a PSD operator, which contributes to the effect of the fourth moment in the excess risk bound. The reader can refer to Appendix F for details of the linear operators.

7.1 Bias-Variance Decomposition

Following the techique used extensively in previous works (Dieuleveut & Bach, 2015; Jain et al., 2018; Zou et al., 2021b; Wu et al., 2022; Liang & Rakhlin, 2020), we decompose the centered iterate $\boldsymbol{\eta}_t$ into the bias sequence $\boldsymbol{\eta}_t^{\text{bias}}$ and the variance sequence $\boldsymbol{\eta}_t^{\text{var}}$, defined recursively as

$$\boldsymbol{\eta}_t^{\text{bias}} = \widehat{\mathbf{A}}_t \boldsymbol{\eta}_{t-1}^{\text{bias}}, \quad \boldsymbol{\eta}_0^{\text{bias}} = \boldsymbol{\eta}_0; \tag{7.1}$$

$$\boldsymbol{\eta}_t^{\text{var}} = \widehat{\mathbf{A}}_t \boldsymbol{\eta}_{t-1}^{\text{var}} + \boldsymbol{\zeta}_t, \quad \boldsymbol{\eta}_0^{\text{var}} = \mathbf{0}. \tag{7.2}$$

The tail averaged iterate is then $\overline{\boldsymbol{\eta}}_{s:s+N} = \overline{\boldsymbol{\eta}}_{s:s+N}^{\text{bias}} + \overline{\boldsymbol{\eta}}_{s:s+N}^{\text{var}}$, where

$$\overline{\boldsymbol{\eta}}_{s:s+N}^{\text{bias}} \coloneqq \frac{1}{N} \sum_{t=s}^{s+N-1} \boldsymbol{\eta}_t^{\text{bias}}, \quad \overline{\boldsymbol{\eta}}_{s:s+N}^{\text{var}} \coloneqq \frac{1}{N} \sum_{t=s}^{s+N-1} \boldsymbol{\eta}_t^{\text{var}}. \tag{7.3}$$

The excess risk can be decomposed into bias and variance:

$$\mathbb{E}[L(\overline{\mathbf{w}}_{s:s+N})] - L(\mathbf{w}^*) = \frac{1}{2} \left\langle \begin{bmatrix} \mathbf{H} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \mathbb{E}[\overline{\boldsymbol{\eta}}_{s:s+N} \otimes \overline{\boldsymbol{\eta}}_{s:s+N}] \right\rangle \le 2 \cdot \text{Bias} + 2 \cdot \text{Variance},$$

where

$$\text{Bias} \coloneqq \frac{1}{2} \left\langle \begin{bmatrix} \mathbf{H} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \mathbb{E}[\overline{\boldsymbol{\eta}}_{s:s+N}^{\text{bias}} \otimes \overline{\boldsymbol{\eta}}_{s:s+N}^{\text{bias}}] \right\rangle, \ \text{Variance} \coloneqq \frac{1}{2} \left\langle \begin{bmatrix} \mathbf{H} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \mathbb{E}[\overline{\boldsymbol{\eta}}_{s:s+N}^{\text{var}} \otimes \overline{\boldsymbol{\eta}}_{s:s+N}^{\text{var}}] \right\rangle.$$

Define the covariance matrices $\mathbf{B}_t \coloneqq \mathbb{E}[\boldsymbol{\eta}_t^{\text{bias}} \otimes \boldsymbol{\eta}_t^{\text{bias}}]$ and $\mathbf{C}_t \coloneqq \mathbb{E}[\boldsymbol{\eta}_t^{\text{var}} \otimes \boldsymbol{\eta}_t^{\text{var}}]$. The recursive forms of $\mathbf{B}_t$ and $\mathbf{C}_t$ then satisfy

$$\mathbf{B}_t = \mathcal{B} \circ \mathbf{B}_{t-1}, \quad \mathbf{B}_0 = \boldsymbol{\eta}_0 \otimes \boldsymbol{\eta}_0; \tag{7.4}$$

$$\mathbf{C}_t = \mathcal{B} \circ \mathbf{C}_{t-1} + \widehat{\boldsymbol{\Sigma}}, \quad \mathbf{C}_0 = \mathbf{0}. \tag{7.5}$$

7.2 Proof of the Bias Bound

In this part, we provide an overview of the analysis of the bias bound in a simplified problem setting. We consider the last bias iterate (i.e., $N = 1$) and assume that $\mathcal{B} = \widetilde{\mathcal{B}}$. The analysis of the general

cases is given in Appendix H. According to the recursive form of $\mathbf{B}_s$ in (7.4), we have $\mathbf{B}_s = \mathcal{B}^s \circ \mathbf{B}_0$. With the assumptions that $\mathcal{B} = \widetilde{\mathcal{B}}$, we have

$$\mathbf{B}_s = \widetilde{\mathcal{B}}^s \circ \mathbf{B}_0 = \mathbf{A}^s \Big(\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \otimes (\mathbf{w}_0 - \mathbf{w}^*)(\mathbf{w}_0 - \mathbf{w}^*)^\top \Big) (\mathbf{A}^s)^\top.$$

Note that $\mathbf{A}$ is block-diagonal with each block being $\mathbf{A}_i$, so bias can be expressed as

$$\text{Bias} = \frac{1}{2} \left\langle \begin{bmatrix} \mathbf{H} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \mathbb{E}[\overline{\boldsymbol{\eta}}^{\text{bias}}_{s:s+N} \otimes \overline{\boldsymbol{\eta}}^{\text{bias}}_{s:s+N}] \right\rangle = \frac{1}{2} \left\langle \begin{bmatrix} \mathbf{H} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \mathbf{B}_s \right\rangle = \frac{1}{2} \sum_{i=1}^{d} \lambda_i w_i^2 \left(\mathbf{A}_i^s \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right)_1^2,$$

where $w_i \coloneqq (\mathbf{w}_0 - \mathbf{w}^*)_i$. The following lemma explicitly characterizes $\mathbf{A}_i^k$:

Lemma 7.1. Let the eigenvalues of $\mathbf{A}_i$ be x_1 and x_2. Then, for any integer $k \geq 1$, we have

$$\mathbf{A}_i^k = \begin{bmatrix} -c(1-\delta\lambda_i) \cdot \frac{x_2^{k-1} - x_1^{k-1}}{x_2 - x_1} & (1-\delta\lambda_i) \cdot \frac{x_2^k - x_1^k}{x_2 - x_1} \\ -c \cdot \frac{x_2^k - x_1^k}{x_2 - x_1} & \frac{x_2^{k+1} - x_1^{k+1}}{x_2 - x_1} \end{bmatrix}.$$

The detailed proof of Lemma 7.1 is given as the proof of Lemma E.3. With Lemma 7.1, we have

$$\mathrm{I} \coloneqq \left(\mathbf{A}_i^s \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right)_1 = (1-\delta\lambda_i) \frac{x_2^{s-1}(x_2 - c) - x_1^{s-1}(x_1 - c)}{x_2 - x_1}.$$

For $i \leq k^\ddagger$ and $i > k^\dagger$, i.e., $\mathbf{A}_i$ has real eigenvalues $x_1 < x_2$, I decays exponentially with the same rate of x_2^s. For $k^\ddagger < i \leq k^\dagger$, i.e., $\mathbf{A}_i$ has complex eigenvalues with $|x_1| = |x_2|$, $|\mathrm{I}|$ is bounded by

$$\begin{aligned} |\mathrm{I}| &= (1-\delta\lambda_i) \left| \frac{x_2^{s-1}(x_2 - c) - x_1^{s-1}(x_1 - c)}{x_2 - x_1} \right| \leq \left| \frac{x_1^{s-1} + x_2^{s-1}}{2} + \frac{x_1 + x_2 - 2c}{2} \cdot \frac{x_2^{s-1} - x_1^{s-1}}{x_2 - x_1} \right| \\ &\leq |x_2|^{s-1} + \frac{|x_1 + x_2 - 2c|}{2} \cdot \left| \frac{x_2^{s-1} - x_1^{s-1}}{x_2 - x_1} \right|, \end{aligned}$$

where the first inequality holds because $0 \leq 1 - \delta\lambda_i \leq 1$, and the second inequality holds due to triangle inequality. For the term $|(x_2^{s-1} - x_1^{s-1})/(x_2 - x_1)|$, note that

$$\left| \frac{x_2^{s-1} - x_1^{s-1}}{x_2 - x_1} \right| = \left| \sum_{k=0}^{s-2} x_2^k x_1^{s-2-k} \right| \leq \sum_{k=0}^{s-2} |x_2^k| \cdot |x_2^{s-2-k}| = \sum_{k=0}^{s-2} |x_2|^k \cdot |x_1|^{s-2-k} = (s-1)|x_2|^{s-2},$$

where the inequality holds due to triangle inequality, and the second inequality holds because $|x_1| = |x_2|$. Therefore, the exponential decay rate of $|\mathrm{I}|$ is $|x_2|^s$. The following lemma provides tight bounds of x_2, thus characterizing the exponential rate of bias decay within each eigen-subspace:

Lemma 7.2. Let x_1, x_2 be the eigenvalues of $\mathbf{A}_i$. Then

(a) When $i \leq k^\ddagger$, $(c\delta - \sqrt{c(q-\delta)(q-c\delta)})/q \leq x_2 \leq c\delta/q$.

(b) When $k^\ddagger < i \leq k^\dagger$, $|x_2| = \sqrt{c(1-\delta\lambda_i)}$.

(c) When $i > k^\dagger$, $1 - (\gamma+\delta)\lambda_i \leq x_2 \leq 1 - (\gamma+\delta)\lambda_i/2$.

The detailed proof of Lemma 7.2 is given in Appendix E.1. We can thus obtain the exponential decay rate of the effective bias.

8 Conclusion

In this work, we consider accelerated SGD with tail averaging for overparameterized linear regression. We provide instance-dependent risk bounds for accelerated SGD that are comprehensively dependent on the spectrum of the data covariance matrix. We show that the variance error of accelerated SGD is always larger than that of SGD. We also show that the bias error of accelerated SGD is smaller than that of SGD along the small eigenvalues subspace but is larger than that of SGD along the small eigenvalues subspace. These together suggest that accelerated SGD outperforms SGD only if the signals mostly align with the small eigenvalues subspaces of the data covariance and that the noise is small. Our results also improve a best-known bound for accelerated SGD in the classic regime (Jain et al., 2018).

ACKNOWLEDGEMENTS

We thank the anonymous reviewers and area chair for their helpful comments. YD and QG are supported in part by the NSF grants IIS-2008981, CHE-2247426, and the Sloan Research Fellowship. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

REFERENCES

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

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

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

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, M. Sham Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Accelerating stochastic gradient descent for least squares regression. *COLT*, pp. 545–604, 2018.

Rahul Kidambi, Praneeth Netrapalli, Prateek Jain, and Sham Kakade. On the insufficiency of existing momentum schemes for stochastic optimization. In *2018 Information Theory and Applications Workshop (ITA)*, pp. 1–9. IEEE, 2018.

Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel "ridgeless" regression can generalize. *The Annals of Statistics*, 48(3):1329–1347, 2020.

Chaoyue Liu and Mikhail Belkin. Accelerating sgd with momentum for over-parameterized learning. *arXiv preprint arXiv:1810.13395*, 2018.

Yu E Nesterov. A method for solving the convex programming problem with convergence rate $o(\frac{1}{k^2})$. In *Dokl. Akad. Nauk SSSR,*, volume 269, pp. 543–547, 1983.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Publishing Company, Incorporated, 1 edition, 2014. ISBN 1461346916.

Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17, 1964.

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

Aditya Varre and Nicolas Flammarion. Accelerated sgd for non-strongly-convex least squares, 2022.

Runzhe Wang, Sadhika Malladi, Tianhao Wang, Kaifeng Lyu, and Zhiyuan Li. The marginal value of momentum for small learning rate sgd. *arXiv preprint arXiv:2307.15196*, 2023.

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

Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, Dean P Foster, and Sham M Kakade. The benefits of implicit regularization from sgd in least squares problems. *The 35th Conference on Neural Information Processing Systems*, 2021a.

Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, and Sham M Kakade. Benign overfitting of constant-stepsize sgd for linear regression. *The 34th Annual Conference on Learning Theory*, 2021b.