Does Momentum Help in Stochastic Optimization? A Sample Complexity Analysis.

Swetha Ganesh *1 **Rohan Deb***†2 **Gugan Thoppe**1 **Amarjit Budhiraja**3

1Department of Computer Science and Automation, Indian Institute of Science, Bengaluru, India
2Department of Computer Science, University of Illinois Urbana-Champaign, USA
3Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, USA

Abstract

Stochastic Heavy Ball (SHB) and Nesterov's Accelerated Stochastic Gradient (ASG) are popular momentum methods in optimization. While the benefits of these acceleration ideas in deterministic settings are well understood, their advantages in stochastic optimization are unclear. Several works have recently claimed that SHB and ASG always help in stochastic optimization. Our work shows that i.) these claims are either flawed or one-sided (e.g., consider only the bias term but not the variance), and ii.) when both these terms are accounted for, SHB and ASG do not always help. Specifically, for *any* quadratic optimization, we obtain a lower bound on the sample complexity of SHB and ASG, accounting for both bias and variance, and show that the vanilla SGD can achieve the same bound.

1 INTRODUCTION

In deterministic convex optimization (when one has access to exact gradients), Gradient Descent (GD) is a popular optimization algorithm [Cauchy, 1847]. In practice, though, exact gradients are not available and one has to rely on their noisy estimates. This brings forth the idea of Stochastic Gradient Descent (SGD). Two classic momentum methods used to accelerate GD are Heavy Ball (HB) [Polyak, 1964, 1987, Qian, 1999] and Nesterov's Accelerated Gradient (NAG) [Nesterov, 1983, 2014, 2005]. Naturally, these momentum-based methods and their variants have also gained significant interest in stochastic settings [Sutskever et al., 2013, Nitanda, 2014a, Hu et al., 2009a]. However, our work shows that the stochastic variants of HB and NAG, i.e., the Stochastic Heavy Ball (SHB) and Nesterov's Accelerated Stochastic Gradient (ASG), are not always better than the vanilla SGD for any quadratic optimization. Specifically, we provide conditions for which the sample complexities of SHB and ASG are never better than that of SGD[1].

We elaborate on the discussion above. The benefit of using momentum in (deterministic) quadratic optimization is the following. Suppose the driving matrix has condition number κ. Then, for any $\epsilon > 0$, GD with an optimal constant stepsize[2] converges to an ϵ-close solution in $\mathcal{O}(\kappa \log \frac{1}{\epsilon})$ iterations. In contrast, both HB and NAG with optimal stepsize and momentum parameters only need $\mathcal{O}(\sqrt{\kappa} \log \frac{1}{\epsilon})$ steps; see, e.g., [Recht, 2010]. Our main claim here is that momentum does not lead to similar advantages in stochastic settings. We use Figure 1 to provide an intuitive justification for this claim. The setup is as follows. We consider a quadratic optimization problem (see Section E for the details) and ensure that only a noisy estimate of its gradient is available in each iteration. This problem is solved using SGD, SHB, and ASG and the three panels show how the Mean Squared Error (MSE) decays for different stepsize and momentum parameter choices. Note that these parameters, once chosen, are fixed, i.e., they do not change from one iteration to the other.

In stochastic settings, the MSE error at any time instance for each of SGD, SHB, and ASG can be broken down into two components: bias and variance. The bias dictates how fast the distance of the initial estimate to the solution is forgotten, while the variance represents a cumulative effect of the noise seen so far. When constant stepsize and momentum parameters are used, the bias decays exponentially fast while the variance converges to some (non-zero) positive constant; this implies the MSE also converges to this constant. Both the rate at which the bias decays and the constant to which the variance converges to are influenced by the stepsize and

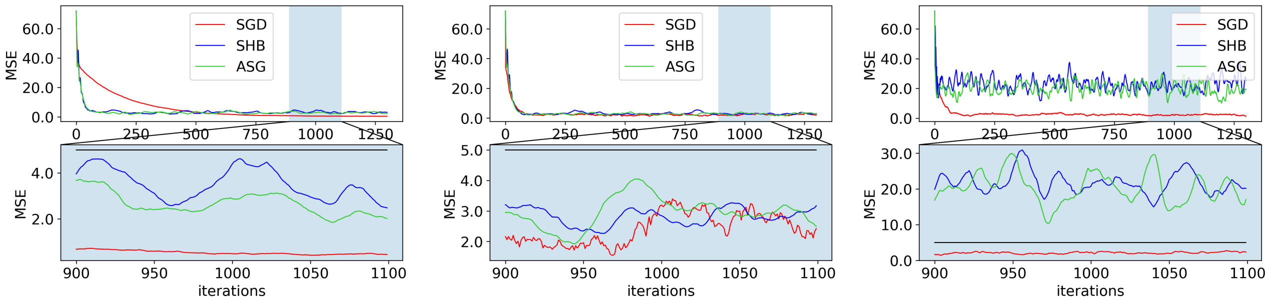
*Equal Contribution

†Part of the research was done while RD was a Project Associate with GT at the Indian Institute of Science, Bangalore.

[1] Sample complexity refers to the number of iterations required to reach an ϵ-ball around the solution. Our statement holds for all sufficiently small ϵ.

[2] Throughout, we only consider algorithms with constant stepsizes, which are widely popular in practice.

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).



(a) Stepsize SGD, SHB, ASG: 0.0025 (b) Stepsize SGD: 0.02; SHB, ASG: 0.0025 (c) Stepsize - SGD, SHB, ASG: 0.02

Figure 1: Comparison of SGD, SHB, and ASG's performances for a 2D quadratic optimization problem (see Section E for details) for the different stepsize choices given above and ϵ-threshold = 5 (denoted by the black horizontal line).

momentum parameter choices.

With the above picture in mind, Figure 1 illustrates how SHB and ASG's performance can be matched by SGD. Figure 1a corresponds to the case where a same stepsize is used in all the three algorithms. In this case, the MSE for the momentum based methods (SHB, ASG) decreases faster initially, but settles at a higher limiting value eventually. Accordingly, one may conjecture that SHB and ASG would have a better sample complexity if the ϵ-threshold for the MSE is set above this limit (one such choice of ϵ in this example is 5). However, Figure 1b shows that SGD enjoys a similar performance for a larger stepsize choice. This time one may conjecture that SHB and ASG's performance can be improved if their stepsizes are also increased similarly. Figure 1c discusses this case when the stepsize for momentum methods is increased to match the new stepsize for SGD. Unfortunately, while MSE for momentum based does decrease faster initially, it also settles at a value that is higher than the threshold that we had set before, i.e., 5.

Related Works: Some recent results [Loizou and Richtárik, 2020, Mou et al., 2020, Assran and Rabbat, 2020, Can et al., 2019] claim that SHB and ASG methods are better than SGD in quadratic or least-squares settings. However, Loizou and Richtárik [2020] needs a strong assumption on noise, which Kidambi et al. [2018, Section 6] claim is information-theoretically impossible even in the simple least squares regression problem. The other results either are based on a one-sided analysis [Can et al., 2019][3] or have a flaw [Mou et al., 2020, Assran and Rabbat, 2020]; see Appendix A.

On the other hand, there are also a few recent negative results on these momentum methods. Devolder et al. [2014] make a similar conclusion to ours in the context of (deterministic) proximal gradient methods and their accelerated variants for smooth convex optimization, when the function can be estimated only up to some (non-random) fixed inaccuracy. Yuan et al. [2016] show that SHB and ASG are equivalent to SGD with a rescaled stepsize. However, this result requires that the stepsize be sufficiently small and the momentum parameter be away from 1. Liu et al. [2021] obtain an expression for the asymptotic variance for SHB and show that it can be matched by that of vanilla SGD with a rescaled stepsize. However, this discussion is only from an asymptotic sense and compares the final size of the ball where the iterates with or without momentum settle, but not the number of iterations needed to reach such a ball. In fact, the asymptotic variance estimate does not provide any information about the sample complexity. In [Kidambi et al., 2018, Liu and Belkin, 2020], for one specific instance of the least squares regression with *vanishing noise*, it is shown that the performance of SHB and ASG cannot be better than that of SGD. Finally, Zhang et al. [2019] consider SHB for quadratic objectives in the noisy setting as our work and provides upper bounds on the rate at which the objective function decreases. They also argue that rescaled SGD performs as well as SHB and demonstrate it empirically but fall short of rigorously coming up with a lower bound that supports their claim.

SHB and ASG have also been studied in the decreasing stepsize setting. Ghadimi et al. [2014] had given the first global convergence of SHB for quadratic objectives while Yang et al. [2016], Promsinchai et al. [2020], Orvieto et al. [2019] gave a.s. convergence rates for convex objectives. In [Sebbouh et al., 2021], improved bounds on both SGD and SHB have been provided, as compared to previously known bounds. Hu et al. [2009b], Ghadimi and Lan [2012], Xiao [2009] study Nesterov's momentum under a decreasing stepsize setting and show that though the momentum scheme accelerates the convergence of the iterates in the initial part, the acceleration is lost in the asymptotic regime. Vaswani et al. [2019] study ASG with a decreasing momentum parameter and show a linear convergence to the optimal point. However, the noise at any stationary point vanishes to zero in their setting. Finally, we also note that other momentum methods have been studied in [Allen-Zhu, 2018, Nitanda, 2014b, Defazio et al., 2014, Johnson and Zhang, 2013, Roux

[3] This work only considers bias, while ignoring variance

et al., 2012] that can provably be shown to have a better performance than SGD.

The current literature can thus be summarized as follows.

Research Gap: Existing works on SHB and ASG fall into two groups: i.) positive - where the results claim advantages of these methods over SGD and ii.) negative - where the results claim the opposite. Results in the positive group either have a one-sided or a flawed analysis, while the ones in the negative apply only in some restricted settings.

Key Contribution: Our work belongs to the negative group: SHB and ASG do not have an advantage over SGD. Specifically, for *all quadratic optimization problems* with persistent noise (noise variance is sufficiently bounded away from zero) and any sufficiently small $\epsilon > 0$, we show that number of iterations needed by SHB and ASG to find an ϵ-optimal solution are not better than that of SGD. More technically, we obtain a lower bound on sample complexities of SHB and ASG (Theorem 2.5) and show that these are of the same order as the corresponding upper bound for SGD (Proposition 2.8). Our proof techniques are also significantly different from those used in existing lower bounds such as [Kidambi et al., 2018, Liu and Belkin, 2020]. This is because, under non-vanishing noise, the expected error contains an additional term that cannot be accounted for from their analyses (see Remark 2.7).

2 MAIN RESULTS

We state our main results here that provide lower and upper bounds on the sample complexities of SHB and ASG. We use these bounds along with those of SGD to show that all these methods need a similar effort to find an ϵ-optimal solution.

Throughout, we consider minimizing

$$f(x) = \frac{1}{2}x^T Ax - b^T x + c, \tag{1}$$

where A is some symmetric $d \times d$ matrix, $b \in \mathbf{R}^d$, and $c \in \mathbf{R}$. The update rules for standard algorithms such as SHB, ASG, and SGD for solving this problem can be jointly expressed as

$$\begin{aligned} x_n &= x_{n-1} + \alpha(b - Ax_{n-1} + M_n) \\ &\quad + \eta(I_d - \alpha\beta A)(x_n - x_{n-1}) \qquad (2) \\ &= x_{n-1} + \alpha(b - A(x_{n-1} + \eta\beta(x_{n-1} - x_{n-2})) + M_n) \\ &\quad + \eta(x_{n-1} - x_{n-2}) \qquad (3) \end{aligned}$$

with $x_{-1} = x_0$. The notation I_d is the $d \times d$ identity matrix, and $M_{n+1} \in \mathbb{R}^d$ is noise. Henceforth, we will refer to the above generic algorithm as Linear Stochastic Approximation with Momentum (LSA-M). Note that LSA-M is equivalent to SGD (if $\eta = 0$ in (2)), to SHB (if $\beta = 0$ in (2)), and to ASG (if $\beta = 1$ in (3)).

We make the following assumption on the driving matrix.

Assumption 2.1 (Driving matrix property). *A is real symmetric and all its eigenvalues are positive.*

We also denote the the eigenvalues of A by $\lambda_{\max} = \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d = \lambda_{\min}$. Under the above assumption, one would expect the iterates in (2) to go to a neighborhood of $x^* := A^{-1}b$.

We next state two assumptions on the noise sequence (M_n), the first is used in Theorem 2.5, while the other is used in Proposition 2.8 and Corollary 2.9. The notation $A \succeq B$ means $A - B$ is positive semi-definite.

Assumption 2.2 (Noise attributes for Theorem 2.5). *(M_n) is a martingale difference sequence with respect to the filtration $(\mathcal{F}_n)$, where $\mathcal{F}_n = \sigma(x_m, M_m; m \leq n)$. Further, $\exists K > 0$ such that $\mathbb{E}[M_{n+1}M_{n+1}^T|\mathcal{F}_n] \succeq KI_d$ a.s. $\forall n \geq 0$.*

Assumption 2.3 (Noise attributes for Proposition 2.8). *(M_n) is a martingale difference sequence with respect to the filtration $(\mathcal{F}_n)$, where $\mathcal{F}_n = \sigma(x_m, M_m; m \leq n)$. Further, $\exists K \geq 0$ such that $\mathbb{E}[\|M_{n+1}\|^2|\mathcal{F}_n] \leq K(1 + \|x_n - x^*\|^2)$ a.s. $\forall n \geq 0$.*

Assumptions 2.2 and 2.3 are standard [Mandt et al., 2017, Jastrzębski et al., 2018, Cheng et al., 2020, Borkar, 2008]. The first of these holds if and only if all the eigenvalues of $\mathbb{E}[M_{n+1}M_{n+1}^T|\mathcal{F}_n]$ are bounded from below by K, i.e., noise is persistent (or non-vanishing) in all directions. On the other hand, Assumption 2.3 requires that the trace of $\mathbb{E}[M_{n+1}M_{n+1}^T|\mathcal{F}_n]$ be bounded from above. This bound can scale with $\|x_n - x^*\|$ and need not vanish near x^*.

Next, we define sample complexity to quantify the effort required by LSA-M to obtain an ϵ-close solution to x^*.

Definition 2.4 (Sample Complexity). *The sample complexity of (2) is the minimum number of iterations n_0 such that the expected error $\mathbb{E}[\|x_n - x^*\|^2] \leq \epsilon$, $\forall n \geq n_0$.*

To enable easy comparison between different algorithms, we shall look at the order of their sample complexities. Towards that, we shall use the notation $n_0 \in \Theta(t)$ to imply that there exist constants c_1 and c_2 (independent of t) such that $c_1 t \leq n_0 \leq c_2 t$. The notation $\tilde{\Theta}(t)$ has a similar meaning but hides the dependence on logarithmic terms. Further, $n_0 \in \Omega(t)$ implies there exists c_1 such that $n_0 \geq c_1 t$ and $n_0 \in \mathcal{O}(t)$ implies there exists c_2 such that $n_0 \leq c_2 t$.

Theorem 2.5. ***(Lower bound on sample complexity).*** *Consider the LSA-M update rule (2), and suppose Assumptions 2.1 and 2.2 hold. Then there exists an $\epsilon' > 0$ such that, for any $\epsilon \in (0, \epsilon')$ and for any choice of $\alpha > 0$, $\beta \in [0, 1]$, and $\eta \in [0, 1]$, the expected error $\mathbb{E}[\|x_{n_0} - x^*\|^2] \geq \epsilon$ for $n_0 \in \tilde{\Theta}\left(\frac{K}{\epsilon\lambda_{\min}^2}\right)$. The constant K here is the one from Assumption 2.2.*

Method	β	η	α
SGD	-	0	$\min\left(\frac{\lambda_{\min}}{\frac{3}{4}\lambda_{\min}^2+C^2K}, \frac{\epsilon\lambda_{\min}}{4C^2K}, \frac{2}{\lambda_{\max}+\lambda_{\min}}\right)$
SHB	0	$\left(1-\frac{\sqrt{\alpha\lambda_{\min}}}{2}\right)^2$	$\min\Big((\frac{\lambda_{\min}^{3/2}}{\frac{3}{8}\lambda_{\min}^2+25C^2K})^2, (\frac{\epsilon(\lambda_{\min})^{3/2}}{200C^2K})^2,$ $(\frac{2}{\sqrt{\lambda_{\min}}+\sqrt{\lambda_{\max}}})^2\Big)$
ASG	1	$\frac{\left(1-\frac{\sqrt{\alpha\lambda_{\min}}}{2}\right)^2}{(1-\alpha\lambda_{\min})}$	$\min\Big((\frac{\lambda_{\min}^{3/2}}{\frac{3}{8}\lambda_{\min}^2+25C^2K})^2, (\frac{\epsilon(\lambda_{\min})^{3/2}}{200C^2K})^2, \frac{1}{\lambda_{\max}}\Big)$

Table 1: Parameter choices for Proposition 2.8. Here $C = 1$ when the matrix A is symmetric and $C = \frac{\sqrt{d}}{\sigma_{\min}(S)\sigma_{\min}(S^{-1})}$ when A is not symmetric, where $\sigma_{\min}(\cdot)$ denotes the smallest singular value and S is the matrix that diagonalizes A, i.e., $S^{-1}AS = D$, a diagonal matrix. When A is symmetric, indeed the three parameter choices correspond to SGD, SHB and ASG. We stick to the same naming convention even when the driving matrix A is not symmetric.

See Section 3 for the proof of the above Theorem.

Remark 2.6. *As stated below* (3)*, LSA-M includes SHB and Nesterov's ASG method as special cases and, hence, the above result directly applies to them. In fact, this is the first lower bound on SHB and ASG's sample complexities in quadratic optimization.*

Remark 2.7. *The lower bounds in Kidambi et al. [2018] and Liu and Belkin [2020] are obtained by viewing the expected error in SHB and ASG iterates for least squares as update rules of the form $z_{n+1} = Pz_n$ for some matrix P [Kidambi et al., 2018, Appendix A, p 16] and [Liu and Belkin, 2020, Appendix C, p 12]). In particular, they obtain bounds on the eigenvalues of P to get the desired claim. In contrast, the error relations for SHB and ASG methods in our setup (quadratic optimization with persistent noise) have the form $z_{n+1} = Pz_n + \alpha W_n$ for some matrix P and vector W_n (cf. 4). This forces us to develop a new proof technique that jointly looks at both these terms and show that at least one of them remains larger than ϵ for the choice of n_0 given in Theorem 2.5.*

We next state our upper bound on the sample complexity of (2) in Proposition 2.8 and Corollary 2.9. Similar bounds already exist in literature when A is assumed to be symmetric and the noise is assumed to be iid with variance bounded by a constant ([Can et al., 2019, Zhang et al., 2019]). Here, we show that a similar upper bound holds under more general settings: i.) A is not symmetric but is diagonalizable and has real positive eigenvalues, and ii.) the noise is a martingale difference sequence satisfying Assumption 2.3.

Proposition 2.8. *Consider the LSA-M update rule* (2)*, and suppose A is a (not necessarily symmetric) real diagonalizable matrix with real positive eigenvalues*[4]*. Further suppose 2.3 holds. Then, $\forall\epsilon > 0$, there exists a choice of α, β and η (see Table 1 for exact values) such that the expected error $\mathbb{E}[\|x_n - x^*\|^2] \leq \epsilon$, $\forall n > n_0$, where*

(i) $n_0 \in \tilde{\Theta}(\frac{1}{\alpha\lambda_{\min}})$, *when* $\eta = 0$, *and*

(ii) $n_0 \in \tilde{\Theta}(\frac{1}{\sqrt{\alpha\lambda_{\min}}})$, *when* $\eta > 0$.

For the proof see Appendix C.

From Table 1, we see that α is a minimum of three terms in each case. The first term arises due to the unbounded noise (Assumption 2.3), the second due to the target neighborhood ϵ and the third from the optimal choice of stepsize in the deterministic (no noise scenario) case. Since the bound on n_0 provided in Proposition 2.8 is in terms of α, the minimum of the three terms dictates the sample complexity. Note that ϵ only influences the middle term in all the choices of α given in Table 1.

Let $\bar{\epsilon} > 0$ be such that, for any $\epsilon \in (0, \bar{\epsilon})$, the value of α equals the middle term in each of the three cases in Table 1. Then the following result is immediate.

Corollary 2.9 (**Upper bound on sample complexity**)**.** *Consider the LSA-M update rule* (2)*, and suppose A is as in Proposition 2.8. Further, suppose Assumption 2.3 holds. Then, for choice of parameters in Table 1, and any $\epsilon \in (0, \bar{\epsilon})$, $\exists n_0 \in \tilde{\Theta}\left(\frac{K}{\epsilon\lambda_{\min}^2}\right)$ such that $\mathbb{E}[\|x_n - x^*\|^2] \leq \epsilon$, $\forall n \geq n_0$. The constant K here is the one from Assumption 2.3.*

Remark 2.10. *From Corollary 2.9, we see that the upper bounds on the sample complexities of SGD, SHB, and*

[4] When A is not symmetric, LSA-M *cannot* be viewed as a gradient-based algorithm for minimizing (1). However, the update rule still makes sense, and it can be seen as one that is useful for solving $Ax = b$.

ASG match the lower bound given in Theorem 2.5 for small enough $\epsilon > 0$. In particular, since an upper bound on the sample complexity of SGD matches a lower bound for SHB and ASG, these latter methods do not always outperform SGD from a sample complexity perspective.

Remark 2.11. *Consider ϵ small enough such that the minimum in choice of α is achieved by the second term in Table 1. For SGD, the stepsize $\alpha \in \Theta(\frac{\epsilon\lambda_{\min}}{K})$ is larger than the choice of stepsize for SHB and ASG, $\alpha \in \Theta(\frac{\epsilon^2\lambda_{\min}^3}{K^2})$. Observe that SGD chooses a larger stepsize than SHB and ASG to reach the ϵ ball. Therefore, although momentum methods appear to have a better performance than SGD if the same stepsize is chosen, SGD can match this performance by re-scaling its stepsize (see Figure 1).*

Remark 2.12. *When the noise is assumed to be bounded by a constant, i.e., $\mathbb{E}[\|M_{n+1}\|^2|\mathcal{F}_n] \leq K$ a.s. in Assumption 2.3, the first term in the choice of α in Table 1 does not appear for all three methods. Under such an assumption, if ϵ is large enough or K is small enough such that the third term in the choice of α is the minimum, then the sample complexity of both SHB and ASG is better than SGD. We emphasize that such improvements are lost when the noise variance is large or the neighbourhood under consideration is small.*

3 PROOF OF THE LOWER BOUND (THEOREM 2.5)

We begin by defining the transformed iterates $\tilde{x}_n = x_n - x^*$ and rewriting (2) as

$$\tilde{X}_n = P\tilde{X}_{n-1} + \alpha W_n, \tag{4}$$

where $\tilde{X}_n \triangleq \begin{pmatrix} \tilde{x}_n \\ \tilde{x}_{n-1} \end{pmatrix}, W_n \triangleq \begin{pmatrix} M_n \\ 0 \end{pmatrix}$ and

$$P \triangleq \begin{pmatrix} I_d - \alpha A + \eta(I_d - \alpha\beta A) & -\eta(I_d - \alpha\beta A) \\ I_d & 0 \end{pmatrix}.$$

We derive the bound in Theorem 2.5 by obtained a lower bound for the error expression $\mathbb{E}[\|\tilde{X}_n\|^2]$.

The proof can be summarized by the following key steps.

1. Transform $\tilde{X}_n$ to obtain $\tilde{Y}_n$ (see (5)). Decompose the 2d-dimensional update rule for $\tilde{Y}_n$ (see (6)) into d separate two-dimensional update rules (see (7)) using a block diagonalization argument.
2. For each of the two-dimensional components of $\tilde{Y}_n$ (denoted $\tilde{Y}_n^{(i)}$, $i = 1, \ldots, d$), obtain a lower bound on the error $\mathbb{E}\|\tilde{Y}_n^{(i)}\|^2$. We do this using the following three steps.
   (a) Decompose the error into two components: one that captures the impact of the initialization (*bias*), and the other that concerns the effect of the cumulative noise (*variance*); see Lemma 3.2.
   (b) Use the above decomposition to derive a lower bound on $\mathbf{E}\|\tilde{Y}_n^{(i)}\|^2$ for the special case of $\beta = 0$. The core idea is to show that the *bias* and the *variance* in $\tilde{Y}_n^{(i)}$ cannot be simultaneously small; see Lemma 3.3.
   (c) Generalize the result to $\beta \in [0, 1]$ case by showing that it can be reduced to the former case.
3. Use the lower bound on $\mathbb{E}\|\tilde{Y}_n^{(i)}\|^2$ from Step 2 to obtain a lower bound on the original error $\mathbb{E}\|\tilde{x}_n\|^2$. This proves the desired result for SHB with $\beta = 0$ and ASG with $\beta = 1$.

Next we describe the technical results involved in each of the above steps.

1. Reducing the $2d$-dimensional updates into d separate two-dimensional updates.

We follow a block diagonalization argument as in [Mou et al., 2020] to transform the update rule (4) below.

Lemma 3.1. *There exists a transformation matrix Z and a block diagonal matrix $B = \text{diag}(B_i)$, where $B_i \in \mathbb{R}^{2\times 2}$, so that*

$$\tilde{Y}_n = Z\tilde{X}_n \qquad \textit{and} \qquad \tilde{W}_n = ZW_n \tag{5}$$

satisfy

$$\tilde{Y}_n = B\tilde{Y}_{n-1} + \alpha\tilde{W}_n. \tag{6}$$

In particular, if we break $\tilde{Y}_n$ into d disjoint components of 2-dimensional vectors, then the i-th component

$$\tilde{Y}_n^{(i)} = \begin{pmatrix} 1 - \alpha\lambda_i + \eta' & -\eta' \\ 1 & 0 \end{pmatrix} \tilde{Y}_{n-1}^{(i)} + \alpha\tilde{W}_n^{(i)} \tag{7}$$

where $\eta' = \eta(1 - \alpha\lambda_i\beta)$.

See Section 3.1 for the proof. Notice that the driving matrix B in the transformed update rule (6) is a block diagonal matrix unlike the driving matrix P in (4). In the next step we exploit this structure to lower bound $\mathbb{E}\|\tilde{Y}_n^{(i)}\|^2$.

2. Bounding the error $\mathbb{E}\|\tilde{Y}_n^{(i)}\|^2$.

We consider the two dimensional decoupled update given in (7) for a specific i and express the lower bound on the sample complexity with respect to λ_i.

(a) Decompose the error $\mathbb{E}\|\tilde{Y}_n^{(i)}\|^2$ as a sum of bias and variance.

First observe that the update from Lemma 3.1 can be re-written as

$$\tilde{Y}_n^{(i)} = B_i^n\tilde{Y}_0^{(i)} + \alpha\sum_{i=0}^{n-1} B_i^{(n-1-i)}\tilde{W}_{i+1}^{(i)}. \tag{8}$$

Taking the square of the norm on both sides of the above equation we get

$$\|\tilde{Y}_n^{(i)}\|^2 = \underbrace{\|B_i^n \tilde{Y}_0^{(i)}\|^2}_{I} + \underbrace{2\alpha \left(B_i^n \tilde{Y}_0^{(i)}\right)^T \left(\sum_{j=0}^{n-1} B_i^{(n-1-i)} \tilde{W}_{j+1}^{(i)}\right)}_{II} + \underbrace{\alpha^2 \left(\sum_{j=0}^{n-1} B_i^{(n-1-i)} \tilde{W}_{j+1}^{(i)}\right)^T \left(\sum_{j=0}^{n-1} B_i^{(n-1-i)} \tilde{W}_{j+1}^{(i)}\right)}_{III}. \tag{9}$$

Using the fact that $(\tilde{W}_n) = (ZW_n)$ is a martingale difference sequence, it can be shown that expectation of term II is 0 and that of term III is $\alpha^2 \mathbb{E}\Big[\sum_{j=0}^{n-1} \|B_i^{n-1-j} \tilde{W}_{j+1}^{(i)}\|^2\Big]$ (See Section 3.2 for details). This leads to the following lemma.

Lemma 3.2. *For the update in* (8) *the error can be decomposed as follows:*

$$\mathbb{E}\|\tilde{Y}_n^{(i)}\|^2 = \underbrace{\|B_i^n \tilde{Y}_0^{(i)}\|^2}_{Bias} + \underbrace{\alpha^2 \mathbb{E}\Big[\sum_{j=0}^{n-1} \|B_i^{n-1-j} \tilde{W}_{j+1}^{(i)}\|^2\Big]}_{Variance}. \tag{10}$$

See Section 3.2 for the proof. The *bias* and *variance* here correspond to that of the i-th block of the transformed iterates in (7).

(b) Bounding the error $\mathbb{E}\|\tilde{Y}_n^{(i)}\|^2$ for $\beta = 0$.
Using the fact that $\eta' = \eta$ when $\beta = 0$, the update in 7 reduces to

$$\tilde{Y}_n^{(i)} = \begin{pmatrix} 1 - \alpha\lambda_i + \eta & -\eta \\ 1 & 0 \end{pmatrix} \tilde{Y}_{n-1}^{(i)} + \alpha \tilde{W}_n^{(i)}.$$

We show that there exists an $\epsilon > 0$ such that for some $n_0 \in \tilde{\Theta}\left(\frac{K}{\epsilon\lambda_i^2}\right)$, either the *bias* or the *variance* is larger than ϵ. This is established in the following key lemma.

Lemma 3.3. *Let* $\epsilon_i' = \min\left(\frac{K}{32\lambda_i^2}, \frac{(\tilde{x}_0^{(i)})^2}{72}\right)$. *Then for any* $\epsilon \in (0, \epsilon_i')$, *and any* $\alpha > 0$, $\beta = 0$, $\eta \in [0,1]$, *there exists* $n_0 \in \tilde{\Theta}\left(\frac{K}{\epsilon\lambda_i^2}\right)$, *such that at least one of the following statements hold:*

(a) $\|B_i^{n_0} \tilde{Y}_0^{(i)}\|^2 > \epsilon$

(b) $\alpha^2 \mathbb{E}\Big[\sum_{j=0}^{n_0-1} \|B_i^{n_0-1-j} \tilde{W}_{j+1}^{(i)}\|^2\Big] > \epsilon.$

See Section 3.3 for the proof. Lemma 3.3 along with Lemma 3.2 immediately provides a lower bound on the error, i.e., $\mathbb{E}\|\tilde{Y}_{n_0}^{(i)}\|^2 > \epsilon$ for $\beta = 0$. Lemma 3.3 is the core of the lower bound analysis and the proof is provided in Section 3.3.

(c) Extending (b) to the case $\beta \in (0,1]$.
We complete Step 2 by extending Lemma 3.3 to the case when $\beta \in [0,1]$ as formalized below.

Lemma 3.4. *Let* ϵ_i' *be defined as in Lemma 3.3. Then for any* $\epsilon \in (0, \epsilon_i')$, *and any* $\alpha > 0$, $\beta = [0,1]$, $\eta \in [0,1]$, *there exists* $n_0 \in \tilde{\Theta}\left(\frac{K}{\epsilon\lambda_i^2}\right)$, *such that at least one of the following statements hold:*

(a) $\|B_i^{n_0} \tilde{Y}_0^{(i)}\|^2 > \epsilon$

(b) $\alpha^2 \mathbb{E}\Big[\sum_{j=0}^{n_0-1} \|B_i^{n_0-1-j} \tilde{W}_{j+1}^{(i)}\|^2\Big] > \epsilon.$

See Section 3.4 for the proof. Note that the general $\beta \in [0,1]$ update rule in (7) is equivalent to the $\beta = 0$ update with η redefined as η' and therefore we can re-use Lemma 3.3 if we can ensure $\eta' \in [0,1]$. We show this holds when $\alpha\lambda_i \leq 1$. For the case $\alpha\lambda_i > 1$, we show that the *variance* term is greater than ϵ thus implying the conclusion of Lemma 3.3.

3. Bounding the original error $\mathbb{E}[\|\tilde{X}_n\|^2]$.
Recall that the original update rule is given by

$$\tilde{X}_n = P\tilde{X}_{n-1} + \alpha W_n.$$

To provide a bound on the error $\mathbb{E}[\|\tilde{X}_n\|^2]$, we invoke Lemma 3.3 for $i = d$ and $\lambda_d = \lambda_{\min}$ and use the fact that Z is an orthogonal matrix. We have

$$\begin{aligned} \mathbb{E}[\|\tilde{X}_{n_0}\|^2] &= \mathbb{E}[\|Z^{-1}\tilde{Y}_{n_0}\|^2] \\ &= \mathbb{E}[\|\tilde{Y}_{n_0}\|^2] \geq \mathbb{E}[\|\tilde{Y}_{n_0}^{(d)}\|^2] \geq \epsilon \end{aligned}$$

for all $\epsilon \in (0, \epsilon_d')$ and for n_0 as defined in Lemma 3.3 with λ_i substituted with $\lambda_{\min}$.
Now to obtain a bound for $\mathbb{E}\|\tilde{x}_n\|^2$ from $\mathbb{E}\|\tilde{X}_n\|^2$, we note that

$$\begin{aligned} 2\max\left(\|\tilde{x}_n\|^2, \|\tilde{x}_{n-1}\|^2\right) &\geq \|\tilde{x}_n\|^2 + \|\tilde{x}_{n-1}\|^2 \\ &= \|\tilde{X}_n\|^2. \end{aligned}$$

Therefore the lower bound on $\mathbb{E}[\|\tilde{X}_n\|^2]$ is enough to prove Theorem 2.5. Choosing $\epsilon' = \epsilon_d'$ and noting that $n_0 \in \tilde{\Theta}\left(\frac{K}{\epsilon\lambda_{\min}^2}\right)$ completes the proof of Theorem 2.5.

3.1 PROOF OF LEMMA 3.1

We first discuss how the update rule for $\tilde{Y}_n$ in (6) can be obtained using that of $\tilde{X}_n$ in (4). Towards this, we define $D = \text{diag}(\lambda_i)_{i=1}^d$. Since A is real symmetric (see Assumption 2.1), it has a spectral decomposition of the form

$A = SDS^{-1}$. We define the transformation matrix Z as

$$Z = E_{2d\times 2d}\begin{pmatrix} S & 0 \\ 0 & S \end{pmatrix} \tag{11}$$

where $E_{2d\times 2d}$ is the permutation matrix that changes the order $(1, 2, \ldots, 2d)$ into $(1, d+1, 2, d+2, \ldots, d, 2d)$.

Since $\tilde{X}_n = P\tilde{X}_{n-1} + \alpha W_n$, we get

$$\begin{aligned}\tilde{Y}_n &= Z\tilde{X}_n = ZP\tilde{X}_{n-1} + \alpha Z W_n \\ &= ZPZ^{-1}\tilde{Y}_{n-1} + \alpha Z W_n = B\tilde{Y}_{n-1} + \alpha Z W_n \\ &= B\tilde{Y}_{n-1} + \alpha \tilde{W}_n,\end{aligned}$$

as desired. The last but one equality follows because $ZPZ^{-1} = B$, which itself holds since

$$\begin{aligned} ZPZ^{-1} &= E_{2d\times 2d}\begin{pmatrix} S & 0 \\ 0 & S \end{pmatrix} P \begin{pmatrix} S^{-1} & 0 \\ 0 & S^{-1} \end{pmatrix} E^{-1}_{2d\times 2d} \\ &\overset{(a)}{=} E_{2d\times 2d}\underbrace{\begin{pmatrix} I_{d\times d} - \alpha D + \eta I_{d\times d} & -\eta I_{d\times d} \\ I_{d\times d} & 0_{d\times d} \end{pmatrix}}_{\Gamma} E_{2d\times 2d} \\ &\overset{(b)}{=} B. \end{aligned}$$

Here (a) follows because $E^{-1}_{2d\times 2d} = E_{2d\times 2d}$. Further (b) follows because the left multiplication of $E_{2d\times 2d}$ to Γ changes the order of rows from $(1, 2, \ldots, 2d)$ to $(1, d+1, 2, d+2, \ldots, d, 2d)$ and the right multiplication of $E_{2d\times 2d}$ changes the order of columns from $(1, 2, \ldots, 2d)$ to $(1, d+1, 2, d+2, \ldots, d, 2d)$ which exactly results in B.

To see why (7) holds, let

$$\tilde{Y}_n = \begin{pmatrix} \tilde{Y}_n^{(1)} \\ \tilde{Y}_n^{(2)} \\ \vdots \\ \tilde{Y}_n^{(d)} \end{pmatrix} \text{ and } \tilde{M}_n = \begin{pmatrix} \tilde{M}_{n,1} \\ \tilde{M}_{n,2} \\ \vdots \\ \tilde{M}_{n,d} \end{pmatrix} = SM_n \text{ , where}$$

$\tilde{Y}_n \in \mathbb{R}^{2d}, \tilde{Y}_n^{(i)} \in \mathbb{R}^2$, $\tilde{M}_n \in \mathbb{R}^d, \tilde{M}_{n,i} \in \mathbb{R}$. Now notice that

$$ZW_n = E_{2d\times 2d}\begin{pmatrix} S & 0 \\ 0 & S \end{pmatrix}\begin{pmatrix} M_n \\ 0 \end{pmatrix} \tag{12}$$

$$= E_{2d\times 2d}\begin{pmatrix} \tilde{M}_n \\ 0 \end{pmatrix} = \begin{pmatrix} \tilde{M}_{n,1} \\ 0 \\ \tilde{M}_{n,2} \\ 0 \\ \vdots \\ \tilde{M}_{n,d} \\ 0 \end{pmatrix},$$

where the last equality follows because the left multiplication of $E_{2d\times 2d}$ changes the order of rows from $(1, 2, \ldots, 2d)$ to $(1, d+1, 2, d+2, \ldots, d, 2d)$. Therefore, $\forall i \in [d]$,

$$\tilde{Y}_n^{(i)} = B_i\tilde{Y}_{n-1}^{(i)} + \alpha\tilde{W}_n^{(i)}$$

where $\tilde{W}_n^{(i)} = \begin{pmatrix} \tilde{M}_{n,i} \\ 0 \end{pmatrix}$.

3.2 PROOF OF LEMMA 3.2

Recall the error expression from (9):

$$\begin{aligned}\|\tilde{Y}_n^{(i)}\|^2 &= \underbrace{\|B_i^n\tilde{Y}_0^{(i)}\|^2}_{I} \\ &\quad + \underbrace{2\alpha\left(B_i^n\tilde{Y}_0^{(i)}\right)^T\left(\sum_{j=0}^{n-1} B_i^{(n-1-i)}\tilde{W}_{j+1}^{(i)}\right)}_{II} \\ &\quad + \underbrace{\alpha^2\left(\sum_{j=0}^{n-1} B_i^{(n-1-i)}\tilde{W}_{j+1}^{(i)}\right)^T\left(\sum_{j=0}^{n-1} B_i^{(n-1-i)}\tilde{W}_{j+1}^{(i)}\right)}_{III}.\end{aligned}$$

Since $\tilde{W}_n = ZW_n$, it follows that $(\tilde{W}_n)$ is also a martingale difference sequence w.r.t. the filtration $(\mathcal{F}_n)$, where $\mathcal{F}_n$ is as in Assumption 2.2. In particular, since $\mathbb{E}[\tilde{W}_n^{(i)}] = 0$ for each n, we get that the expectation of Term II is 0. With regards to Term III, we have

$$\begin{aligned}&\alpha^2\left(\sum_{j=0}^{n-1} B_i^{(n-1-j)}\tilde{W}_{j+1}^{(i)}\right)^T\left(\sum_{j=0}^{n-1} B_i^{(n-1-j)}\tilde{W}_{j+1}^{(i)}\right) \\ &= \alpha^2\sum_{j,k}(\tilde{W}_{j+1}^{(i)})^T(B_i^{(n-1-j)})^T B_i^{(n-1-k)}\tilde{W}_{k+1}^{(k)} \\ &= \underbrace{\alpha^2\sum_{j\neq k}(\tilde{W}_{j+1}^{(i)})^T(B_i^{(n-1-j)})^T B_i^{(n-1-k)}\tilde{W}_{k+1}^{(k)}}_{III(a)} \\ &\quad + \underbrace{\alpha^2\sum_{j}\|B_i^{(n-1-j)}\tilde{W}_{j+1}^{(i)}\|^2}_{III(b)}\end{aligned}$$

We now show that the expectation of $III(a)$ is 0. Without loss of generality, suppose $j < k$. Then,

$$\begin{aligned}&\mathbb{E}\left[(\tilde{W}_{j+1}^{(i)})^T(B_i^{(n-1-j)})^T B_i^{(n-1-k)}\tilde{W}_{k+1}^{(i)}\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[(\tilde{W}_{i+1}^{(i)})^T(B_i^{(n-1-i)})^T B_i^{(n-1-j)}\tilde{W}_{j+1}^{(i)}|\mathcal{F}_j\right]\right] \\ &= \mathbb{E}\left[(\tilde{W}_{i+1}^{(i)})^T(B_i^{(n-1-i)})^T B_i^{(n-1-j)}\mathbb{E}[\tilde{W}_{j+1}^{(i)}|\mathcal{F}_j]\right] = 0.\end{aligned}$$

Therefore, taking expectation on both sides of (9) gives

$$\begin{aligned}\mathbb{E}\|\tilde{Y}_n^{(i)}\|^2 &= \underbrace{\|B_i^n\tilde{X}_0\|^2}_{I} \\ &\quad + \underbrace{\mathbb{E}\Big[\alpha^2\sum_{j=0}^{n-1}\|B_i^{(n-1-j)}\tilde{W}_{j+1}^{(i)}\|^2\Big]}_{III(b)}\end{aligned} \tag{13}$$

3.3 PROOF OF LEMMA 3.3

This is the key result in the lower bound proof. Here we outline the main steps involved in proving the result. The detailed proofs of the all auxiliary lemmas are pushed to Appendix B.

Before we proceed with the main proof, we provide a lower bound on the *variance* term in the following lemma.

Lemma 3.5. *Under Assumption 2.2 and n_0 as in Lemma 3.3, the variance term in* (10) *can be lower bounded as follows:*

$$\alpha^2 \mathbb{E}\left[\sum_{j=0}^{n_0-1} \|B_i^{n_0-1-j}\tilde{W}_{j+1}^{(i)}\|^2\right] \geq \alpha^2 K \sum_{j=0}^{n_0-1} \|B_i^j e_1\|^2$$

where $e_1 = \begin{pmatrix}1\\0\end{pmatrix}$ *and K is as in Assumption 2.2.*

For convenience we redefine the term in the right hand side of the above inequality as the *variance*. If α and η are such that $\|B_i^{n_0}\tilde{Y}_0^{(i)}\|^2 > \epsilon$, then Lemma 3.3 immediately follows for this choice of α and η. We now consider the case where α and η are such that $\|B_i^{n_0}\tilde{Y}_0^{(i)}\|^2 \leq \epsilon$. Now we show that for this choice of α and η, the *variance* is necessarily greater than ϵ. Let $\mu_+^{(i)}$ and $\mu_-^{(i)}$ be the eigenvalues of B_i. It is easy to check that

$$\begin{aligned} \mu_+^{(i)} &= \frac{1}{2}\left((1-\alpha\lambda_i+\eta)+\Delta^{(i)}\right) \\ \mu_-^{(i)} &= \frac{1}{2}\left((1-\alpha\lambda_i+\eta)-\Delta^{(i)}\right) \end{aligned} \tag{14}$$

where $\Delta^{(i)} = \sqrt{(1-\alpha\lambda_i+\eta)^2 - 4\eta}$.

Recall that $\epsilon \in (0, \epsilon_i')$ in Lemma 3.3 and therefore $\|B_i^{n_0}\tilde{Y}_0^{(i)}\|^2 \leq \epsilon$ implies $\|B_i^{n_0}\tilde{Y}_0^{(i)}\|^2 < \epsilon_i'$. The following Lemma provides a lower bound on the *variance* in terms of the eigen values of B_i and the momentum parameter η assuming the *bias* is less than ϵ_i'.

Lemma 3.6. *Let $\alpha > 0$ and $\eta \in [0,1]$ such that $\|B_i^{n_0}\tilde{X}_0\|^2 < \epsilon_i'$. Then*

$$\alpha^2 K \sum_{j=0}^{n_0-1} \|B_i^j e_1\|^2 \geq \frac{\alpha^2 K}{2(1-\mu_+^2)(1-\mu_-^2)(1-\eta)}.$$

It can be shown that $(1-\mu_+^2)(1-\mu_-^2) = \alpha\lambda_i$ and therefore the RHS in the above expression reduces to $\frac{\alpha K}{2\lambda_i(1-\eta)}$. We define the following function

$$Q(\eta;\alpha,\lambda_i) \equiv \frac{\alpha K}{2\lambda_i(1-\eta)} \frac{1}{(1-\rho(B_i))}$$

where $\rho(B_i) = |\mu_+^{(i)}|$ is the spectral radius of B_i. Note that $\rho(B_i)$ depends on η (see 14). Now to obtain a further lower bound on the *variance* we optimize over the choice of η and show that

$$Q(\eta;\alpha,\lambda_i) \geq \frac{K}{16\lambda_i^2}$$

Combining this with the definition of Q and Lemma 3.6 gives the following bound:

$$\alpha^2 K \sum_{j=0}^{n_0-1} \left\|B_i^j e_1\right\|^2 \geq \frac{K}{16\lambda_i^2}(1-\rho(B_i))$$

The following lemma proves all these above claims.

Lemma 3.7. *Let $\alpha > 0$ and $\eta \in [0,1]$ such that $\|B_i^{n_0}\tilde{X}_0\|^2 < \epsilon_i'$. Then we have the following bound $\alpha^2 K \sum_{j=0}^{n_0-1} \|B_i^j e_1\|^2 \geq \frac{K}{16\lambda_i^2}(1-\rho(B_i))$.*

Lastly, to show that the *variance* is lower bounded by $\epsilon \in (0, \epsilon')$, we need to show that $(1-\rho(B_i)) \geq \frac{16\lambda_i^2}{K}\epsilon$. The choice of n_0 and the fact that we assumed $\|B_i^{n_0}\tilde{Y}_0^{(i)}\| < \epsilon$ exactly ensures that. The following lemma proves this claim.

Lemma 3.8. *For any $\epsilon \in (0, \epsilon_i')$, if $\|B_i^{n_0}\tilde{Y}_0^{(i)}\| < \epsilon$, then $1-\rho(B_i) \geq \frac{16\lambda_i^2}{K}\epsilon$.*

This completes the proof of Lemma 3.3.

3.4 PROOF OF LEMMA 3.4

We handle the cases $\alpha\lambda_i \leq 1$ and $\alpha\lambda_i > 1$ separately.

Case 1 ($\boldsymbol{\alpha\lambda_i \leq 1}$): Observe that the general β update rule in (7) is equivalent to the $\beta = 0$ update with η redefined as η'. Moreover in this case $\eta' \in [0,1]$. To see this first observe that

$$\eta' = \eta(1-\alpha\lambda_i\beta) \geq \eta(1-\beta) \geq 0.$$

Here the first inequality follows because $\alpha\lambda_i \leq 1$ and the second inequality follows because $\beta, \eta \in [0,1]$.

Therefore in this case Lemma 3.3 holds with η redefined as η'.

Case 2 ($\boldsymbol{\alpha\lambda_i > 1}$): In this case we show that the variance term is greater than ϵ. This follows as shown below

$$\begin{aligned} \alpha^2 \mathbb{E}\left[\sum_{j=0}^{n_0-1} \|B_i^{n_0-1-j}\tilde{W}_{j+1}^{(i)}\|^2\right] &\overset{(A)}{\geq} \alpha^2 K \sum_{j=0}^{n_0-1} \|B_i^j e_1\|^2 \\ &\overset{(B)}{\geq} \alpha^2 K \overset{(C)}{>} \frac{K}{\lambda_i^2} \overset{(D)}{>} \epsilon. \end{aligned}$$

Here (A) follows from Lemma 3.5, (B) follows from non-negativity of norm and lower bounding the sum with the $j = 0$ term and (C) follows since $\alpha\lambda_i > 1$. Finally (D) follows for any $\epsilon < \frac{K}{\lambda_i^2}$ which in turn is smaller than ϵ_i' as defined in Lemma 3.3.

4 CONCLUDING REMARKS

In this work, we analyze the sample complexity of SHB and ASG and provide matching lower and upper bounds up to constants and logarithmic terms. More importantly, we show that the same sample complexity bound can be obtained by standard SGD. Our work also calls into question some of the recent positive results in favour of SHB and ASG in the stochastic regime. We show that such improvements do not take into account all the terms involved in the error decomposition, or have major flaws. We emphasize that our results hold specifically for SHB and ASG. Other momentum methods could offer provable improvements over SGD [Jain et al., 2018, Liu and Belkin, 2020].

ACKNOWLEDGEMENTS

SG was supported by the Prime Minister's Research Fellowship (PMRF). RD was supported in part by grants from the National Science Foundation (NSF) through awards IIS 21-31335, OAC 21-30835, DBI 20-21898, as well as a C3.ai research award. GT was supported in part by DST-SERB's Core Research Grant CRG/2021/00833, in part by IISc Start-up grants SG/MHRD-19-0054 and SR/MHRD-19-0040, and in part by the "Pratiksha Trust Young Investigator" award. AB was supported in part by the NSF (DMS-2152577, DMS-2134107).

References

Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods, 2018.

M. Assran and M. Rabbat. On the convergence of nesterov's accelerated gradient method in stochastic settings. *Proceedings of the 37th International Conference on Machine Learning, PMLR*, 119:410–420, 2020.

V. S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, 2008. ISBN 9780521515924. URL https://books.google.co.in/books?id=QLxIvgAACAAJ.

Bugra Can, Mert Gurbuzbalaban, and Lingjiong Zhu. Accelerated linear convergence of stochastic momentum methods in Wasserstein distances. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 891–901. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/can19a.html.

L. A. Cauchy. Methode generale pour la resolution des systemes d'equations simultanees. *C.R. Acad. Sci. Paris*, 25: 536–538, 1847. URL https://cir.nii.ac.jp/crid/1573387450834953216.

Xiang Cheng, Dong Yin, Peter Bartlett, and Michael Jordan. Stochastic gradient and Langevin processes. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1810–1819. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/cheng20e.html.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper/2014/file/ede7e2b6d13a41ddf9f4bdef84fdc737-Paper.pdf.

Olivier Devolder, Franccois Glineur, and Yurii Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146:37–75, 2014.

Euhanna Ghadimi, Hamid Reza Feyzmahdavian, and Mikael Johansson. Global convergence of the heavy-ball method for convex optimization, 2014. URL https://arxiv.org/abs/1412.7457.

Saeed Ghadimi and Guanghui Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012. doi: 10.1137/110848864. URL https://doi.org/10.1137/110848864.

Chonghai Hu, Weike Pan, and James Kwok. Accelerated gradient methods for stochastic optimization and online learning. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009a. URL https://proceedings.neurips.cc/paper/2009/file/ec5aa0b7846082a2415f0902f0da88f2-Paper.pdf.

Chonghai Hu, Weike Pan, and James Kwok. Accelerated gradient methods for stochastic optimization and online learning. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009b. URL https://proceedings.neurips.cc/paper/2009/file/ec5aa0b7846082a2415f0902f0da88f2-Paper.pdf.

Prateek Jain, Sham M. Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Accelerating stochastic gradient descent for least squares regression. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 545–604. PMLR, 06–09 Jul 2018. URL https://proceedings.mlr.press/v75/jain18a.html.

Stanislaw Jastrzębski, Zac Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Amos Storkey, and Yoshua Bengio. Three factors influencing minima in SGD, 2018. URL https://openreview.net/forum?id=rJma2bZCW.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper/2013/file/ac1dd209cbcc5e5d1c6e28598e8cbbe8-Paper.pdf.

Rahul Kidambi, Praneeth Netrapalli, Prateek Jain, and Sham M. Kakade. On the insufficiency of existing momentum schemes for stochastic optimization. *CoRR*, abs/1803.05591, 2018. URL http://arxiv.org/abs/1803.05591.

Chaoyue Liu and Mikhail Belkin. Accelerating sgd with momentum for over-parameterized learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1gixp4FPH.

Kangqiao Liu, Liu Ziyin, and Masahito Ueda. Noise and fluctuation of finite learning rate stochastic gradient descent. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7045–7056. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/liu21ad.html.

N. Loizou and P. Richtárik. Momentum and stochastic momentum for stochastic gradient, newton, proximal point and subspace descent methods. *Computational Optimization and Applications*, 77:653–710, 2020.

Stephan Mandt, Matthew D Hoffman, and David M Blei. Stochastic gradient descent as approximate bayesian inference. *Journal of Machine Learning Research*, 18:1–35, 2017.

Wenlong Mou, Chris Junchi Li, Martin J Wainwright, Peter L Bartlett, and Michael I Jordan. On linear stochastic approximation: Fine-grained Polyak-Ruppert and non-asymptotic concentration. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2947–2997. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/mou20a.html.

Y. Nesterov. A method of solving a convex programming problem with convergence rate $o\left(\frac{1}{k^2}\right)$. *Soviet Mathematics Doklady*, 269:543–547, 1983.

Yu Nesterov. Smooth minimization of non-smooth functions. *Math. Program.*, 103(1):127–152, may 2005. ISSN 0025-5610. doi: 10.1007/s10107-004-0552-5. URL https://doi.org/10.1007/s10107-004-0552-5.

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

Atsushi Nitanda. Stochastic proximal gradient descent with acceleration techniques. In *NIPS*, 2014a.

Atsushi Nitanda. Stochastic proximal gradient descent with acceleration techniques. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014b. URL https://proceedings.neurips.cc/paper/2014/file/d554f7bb7be44a7267068a7df88ddd20-Paper.pdf.

Antonio Orvieto, Jonas Kohler, and Aurelien Lucchi. The role of memory in stochastic optimization, 2019. URL https://arxiv.org/abs/1907.01678.

B. Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr Computational Mathematics and Mathematical Physics*, 4:1–17, 12 1964. doi: 10.1016/0041-5553(64)90137-5.

Boris Polyak. *Introduction to Optimization*. Optimization Software, NY, 07 1987.

Porntip Promsinchai, A. Farajzadeh, and Narin Petrot. Stochastic heavy-ball method for constrained stochastic optimization problems. *Acta Mathematica Vietnamica*, 45, 01 2020. doi: 10.1007/s40306-019-00357-y.

Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural Netw.*, 12(1):145–151, jan 1999. ISSN 0893-6080. doi: 10.1016/S0893-6080(98)00116-6. URL https://doi.org/10.1016/S0893-6080(98)00116-6.

Benjamin Recht. Cs726-lyapunov analysis and the heavy ball method. *Department of Computer Sciences, University of Wisconsin–Madison*, 2010.

Nicolas Le Roux, Mark Schmidt, and Francis Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'12, page 2663–2671, Red Hook, NY, USA, 2012. Curran Associates Inc.

Othmane Sebbouh, Robert M Gower, and Aaron Defazio. Almost sure convergence rates for stochastic gradient descent and stochastic heavy ball. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 3935–3971. PMLR, 15–19 Aug 2021. URL https://proceedings.mlr.press/v134/sebbouh21a.html.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1139–1147, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/sutskever13.html.

Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1195–1204. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/vaswani19a.html.

Lin Xiao. Dual averaging method for regularized stochastic learning and online optimization. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009. URL https://proceedings.neurips.cc/paper/2009/file/7cce53cf90577442771720a370c3c723-Paper.pdf.

Tianbao Yang, Qihang Lin, and Zhe Li. Unified convergence analysis of stochastic momentum methods for convex and non-convex optimization, 2016. URL https://arxiv.org/abs/1604.03257.

Kun Yuan, Bicheng Ying, and Ali H. Sayed. On the influence of momentum acceleration on online learning. *Journal of Machine Learning Research*, 17(192):1–66, 2016. URL http://jmlr.org/papers/v17/16-157.html.

Guodong Zhang, Lala Li, Zachary Nado, James Martens, Sushant Sachdeva, George Dahl, Chris Shallue, and Roger B Grosse. Which algorithmic choices matter at which batch sizes? insights from a noisy quadratic model. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/e0eacd983971634327ae1819ea8b6214-Paper.pdf.