2024 American Control Conference (ACC)
July 8-12, 2024. Toronto, Canada

Tradeoffs between convergence speed and noise amplification in first-order optimization: the role of averaging

Samantha Samuelson and Mihailo R. Jovanović

Abstract— **We study the effect of averaging on convergence speed and noise amplification of the heavy-ball algorithm in the presence of additive white stochastic disturbances. For strongly convex quadratic problems, we show that averaging over the entire algorithmic history eliminates steady-state variance of the averaged output at the expense of slowing down convergence to a sub-linear rate. In contrast, finite window averaging converges with a linear rate but it leads to a non-zero value of the steady-state variance. While this value is smaller than the steady-state variance of the iterates of the heavy-ball algorithm, it has the same orderwise dependence on the condition number. We also show that the finite window averaging increases the upper bound on the expected error at iteration t by a constant factor that depends on the length of the averaging window.**

Index Terms— **Convex optimization, gradient descent, heavy-ball method, noisy gradients, performance trade-offs.**

I. Introduction

Accelerated first-order optimization algorithms [1]–[4] are widely used in a variety of optimization applications because of their desirable asymptotic behavior and low per-iteration complexity [5]–[9]. Favorable convergence behavior of accelerated algorithms relative to gradient descent comes at the expense of compromised transient responses [10] and increased sensitivity to gradient noise [11]–[17]. Both of these exhibit undesirable scaling with the condition number [18]–[20].

A growing body of work [21]–[24] considers performance of accelerated methods in a stochastic setting. One of the simplest approaches to mitigating challenges associated with uncertainty in gradient evaluation is to average the algorithmic iterates over time. Intuitively, the average output is expected to produce a smaller steady-state variance and the expected value of the error in the averaged output at a given iteration t should be larger than the error resulting from non-averaged algorithmic iterates. In this paper, we quantify the effects of averaging on convergence properties and variance amplification of the heavy-ball algorithm applied to strongly convex quadratic problems.

Recent work [25] shows that averaging over the entire algorithmic history offers a reduction of worst-case transient growth for the heavy-ball algorithm. We additionally consider the impact of averaging over a moving window of a fixed length, but focus on the asymptotic behavior of the expected value and variance of the averaged output. We show that while averaging over the entire algorithmic history eliminates steady state variance, it reduces the convergence speed to a sub-linear rate. In contrast, averaging over a moving window of fixed length d offers some reduction in steady-state variance while maintaining linear convergence rate.

Our presentation is organized as follows. In Section II, we motivate the problem and provide background material. In Section III, we summarize our main results that quantify the influence of averaging on the trade-off between convergence speed and noise amplification for the heavy-ball method applied to strongly convex quadratic problems. In Section IV, we provide an example that demonstrates the merits and the effectiveness of different averaging schemes on convergence properties and variance amplification for the heavy-ball method. In Section V, we conclude our presentation. The proofs of technical results are relegated to the appendix.

Notation: We use Θ to indicate relative scaling of asymptotic behavior; $g(x) = \Theta(f(x))$ if there exist positive constants c_1, c_2 and x_0 such that the functions g and $f\colon \mathbb{R} \to \mathbb{R}$ satisfy $0 \leq c_1 f(x) \leq g(x) \leq c_2 f(x)$ for all $x \geq x_0$. We use $o(\cdot)$ to indicate relative speed of asymptotic behavior; $g(x) = o(f(x))$ if the functions g and $f\colon \mathbb{R} \to \mathbb{R}$ satisfy $\lim_{t\to\infty} g(x)/f(x) = 0$, i.e., $g(x)$ tends to zero faster than $f(x)$.

II. Motivation and background

Our aim is to study the influence of averaging on convergence rate and noise amplification of first-order algorithms for unconstrained optimization problems

$$\underset{x}{\text{minimize}} \;\; f(x) \tag{1}$$

where $x \in \mathbb{R}^n$ is the optimization variable and $f\colon \mathbb{R}^n \to \mathbb{R}$ is the objective function. In particular, we confine our attention to the heavy-ball algorithm

$$x^{t+2} = (1+\beta)x^{t+1} - \beta x^t - \alpha \nabla f\big(x^{t+1}\big) + w^t \tag{2}$$

where t is the iteration index, α is the stepsize, β is the momentum parameter, and w^t is a white noise disturbance that can account for uncertainty in gradient evaluation, with

$$\mathbb{E}[w^t] = 0, \;\; \mathbb{E}[w^t (w^\tau)^T] = I\delta(t-\tau). \tag{3}$$

Building on the recent work which focuses on the influence of algorithmic parameters on the trade-off between convergence rate and noise amplification [18]–[20], in this paper we examine the impact of averaging on this trade-off. In particular, we average x^t over a moving window of a fixed

The authors are with the Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA 90089. E-mails: ({sasamuel, mihailo}@usc.edu).

length $d \in \mathbb{N}$,

$$z_d^t \;=\; \frac{1}{d} \sum_{k\,=\,t\,-\,d\,+\,1}^{t} x^k. \tag{4a}$$

For $d = 1$, the averaging does not take place, i.e., $x_1^t = x^t$; and, for $d = t$, the averaging is done over the entire algorithmic history,

$$z_t^t \;=\; \frac{1}{t} \sum_{k\,=\,1}^{t} x^k \tag{4b}$$

For iterations with $t < d$, the definition of z_d^t in (4a) is not applicable and we define $z_d^t = z_t^t$ when $t < d$.

We study the role of averaging for the class $\mathcal{Q}_m^L$ of m-strongly convex L-smooth quadratic objective functions,

$$f(x) \;=\; \frac{1}{2}\, x^T Q x \;-\; q^T x \tag{5}$$

where m and L are parameters of strong convexity and Lipschitz continuity, $Q \in \mathbb{R}^{n\times n}$ is the symmetric positive definite Hessian matrix, $mI \preceq Q \preceq LI$, and $\kappa := L/m$ is the condition number determined by the ratio of the largest and the smallest eigenvalues of the matrix Q.

In what follows, without loss of generality we set $q = 0$ in (5) so that the optimal solution to (1) with strongly convex quadratic objective function (5) is $x^\star = Q^{-1}q = 0$.

A. Modal decomposition

For quadratic objective function (5), the gradient $\nabla f(x) = Qx - q$ is an affine function of x and (2) with constant algorithmic parameters admits an LTI state-space representation,

$$\begin{aligned} \psi^{t+1} \;&=\; A\psi^t \;+\; Bw^t \\ y^t \;&=\; C\psi^t. \end{aligned} \tag{6a}$$

Here, $y^t := x^t - x^\star = x^t$ is the distance to the optimal solution $x^\star = Q^{-1}q = 0$, ψ^t is the state vector,

$$\psi^t \;=\; \begin{bmatrix} (y^t)^T & (y^{t+1})^T \end{bmatrix}^T \tag{6b}$$

and A, B, C are constant matrices determined by

$$\begin{aligned} A \;&=\; \begin{bmatrix} 0 & I \\ -\beta I & (1+\beta)I - \alpha Q \end{bmatrix} \\ B \;&=\; \begin{bmatrix} 0 & I \end{bmatrix}^T,\; C \;=\; \begin{bmatrix} I & 0 \end{bmatrix} \end{aligned} \tag{6c}$$

The eigenvalue decomposition of the Hessian matrix, $Q = V\Lambda V^T$, can be used to bring matrices in (6) into their block diagonal forms. Here, V is an orthogonal matrix of the eigenvectors of Q, Λ is a diagonal matrix of the corresponding eigenvalues, and the change of variables,

$$\hat{x} \;:=\; V^T x,\;\; \hat{w} \;:=\; V^T w \tag{7}$$

allows us to transform system (6) into a family of n decoupled subsystems parameterized by the ith eigenvalue λ_i of the Hessian matrix $Q \in \mathbb{R}^{n\times n}$,

$$\begin{aligned} \hat{\psi}_i^{t+1} \;&=\; \hat{A}(\lambda_i)\hat{\psi}_i^t \;+\; \hat{B}\hat{w}_i^t \\ \hat{y}_i^t \;&=\; \hat{C}\hat{\psi}_i^t. \end{aligned} \tag{8a}$$

The ith component of the vector $\hat{w}$ is given by $\hat{w}_i$ and

$$\begin{aligned} \hat{A}(\lambda_i) \;&=\; \begin{bmatrix} 0 & 1 \\ -\beta & (1+\beta) - \alpha\lambda_i \end{bmatrix} \\ \hat{B} \;&=\; \begin{bmatrix} 0 & 1 \end{bmatrix}^T,\;\; \hat{C} \;=\; \begin{bmatrix} 1 & 0 \end{bmatrix}. \end{aligned} \tag{8b}$$

B. Convergence rate

System (6) is stable if absolute values of the eigenvalues of matrices $\hat{A}(\lambda_i)$ are less than one for each $i = 1, \ldots, n$. We recall that stability of (6) implies that, in the absence of white noise, iterations of the algorithm satisfy

$$\|\psi^t\| \;\le\; c\rho^t \|\psi^0\| \tag{9}$$

for some positive constant c, where the convergence rate is given by $\rho = \max_{\lambda\in[m,L]} |\mathrm{eig}(\hat{A}(\lambda))|$.

In the presence of white noise, the expected value of the state vector ψ^t is governed by $\mathbb{E}(\psi^{t+1}) = A\,\mathbb{E}(\psi^t)$. Thus, $\mathbb{E}(\psi^{t+1}) = A^t\,\mathbb{E}(\psi^0)$ and

$$\left\|\mathbb{E}\left(\psi^t\right)\right\| \;\le\; c\rho^t \left\|\mathbb{E}\left(\psi^0\right)\right\|. \tag{10}$$

The best converge rate for strongly convex quadratic problems is achieved by the heavy-ball method [5],

$$\rho \;=\; 1 \;-\; \frac{2}{\sqrt{\kappa} \;+\; 1} \tag{11a}$$

with the following values of the algorithmic parameters

$$\alpha \;=\; \frac{4}{(\sqrt{L} \;+\; \sqrt{m})^2},\; \beta \;=\; \left(1 \;-\; \frac{2}{\sqrt{\kappa} \;+\; 1}\right)^2 \;=\; \rho^2. \tag{11b}$$

This improves the optimal rate achieved by gradient descent

$$\rho \;=\; 1 \;-\; \frac{2}{\kappa \;+\; 1}$$

with the following value of the stepsize $\alpha = 2/(L + m)$.

Our objective is to quantify the convergence rate of the expected value of the averaged output (4) for the heavy-ball method with the optimal values of parameters (11b).

C. Noise amplification

The variance of the error in the optimization variable $y^t := x^t - x^\star = x^t$ is determined by

$$J^t \;:=\; \mathbb{E}\left[\|y^t\|^2\right] \;=\; \sum_{i=1}^{n} \hat{J}^t(\lambda_i) \tag{12}$$

where $\hat{J}^t(\lambda_i) = \mathbb{E}[\|\hat{y}_i^t\|^2]$ denotes the variance amplification of the ith subsystem (8). In particular, the algebraic Lyapunov equation

$$\hat{P}_i^{t+1} \;=\; \hat{A}(\lambda_i)\hat{P}_i^t\hat{A}^T(\lambda_i) + \hat{B}\hat{B}^T \tag{13}$$

can be used to compute the modal contribution of the ith eigenvalue λ_i of Q to the variance amplification as

$$\hat{J}^t(\lambda_i) \;=\; \mathrm{trace}\left(\hat{C}\hat{P}_i^t\hat{C}^T\right)$$

where $\hat{P}_i^t$ denotes the covariance matrix of the state vector $\hat{\psi}_i^t$, $\hat{P}_i^t := \mathbb{E}[\hat{\psi}_i^t(\hat{\psi}_i^t)^T]$. For stable systems, $\hat{P}_i^t$ approaches

its steady-state value $\hat{P}_i^\infty$ as t goes to infinity,

$$\hat{P}_i^\infty \;=\; \hat{A}(\lambda_i)\hat{P}_i^\infty \hat{A}^T(\lambda_i) + \hat{B}\hat{B}^T \tag{14}$$

and the steady-state variance is determined by $\hat{J}^\infty(\lambda_i) = \mathrm{trace}\,(\hat{C}\hat{P}_i^\infty \hat{C}^T)$.

Favorable convergence properties of the heavy-ball algorithm relative to gradient descent come at the expense of compromising dependence of the steady-state variance on the condition number [18]–[20],

$$J_{HB}^\infty \;=\; \Theta\left(\kappa^{3/2}\right), \quad J_{GD}^\infty \;=\; \Theta(\kappa).$$

In this paper, we investigate the impact of averaging on the trade-off between convergence speed and noise amplification.

III. Main results

In this section, we summarize our main results that quantify the influence of averaging on the trade-off between convergence speed and noise amplification for the heavy-ball method. For strongly convex quadratic problems, we show that averaging over the entire algorithmic history eliminates steady-state variance of the averaged output at the expense of slowing down convergence to a sub-linear rate. In contrast, finite window averaging converges with a linear rate but it leads to a non-zero value of the steady-state variance. While this value is smaller than the steady-state variance of the algorithmic iterates, it has the same orderwise dependence on the condition number κ. We also show that the finite window averaging increases the upper bound on the expected error at iteration t by a constant factor that depends on the length of the averaging window.

We first quantify the convergence rate of the expected error of the averaged output z_d^t given by (4), for both fixed window and entire history averaging schemes, and then describe the influence of averaging on the variance.

Theorem 1: Let the heavy-ball algorithm (2) with the strongly convex quadratic objective function f converge linearly with rate ρ,

$$\left\|\mathbb{E}\left(x^t\right)\right\| \;\leq\; c\,\rho^t\,\mathbb{E}\left(\|\psi_0\|\right). \tag{15a}$$

Then the expected error $\mathbb{E}\left(z_d^t\right)$ converges linearly with rate ρ when the averaging window length $d \geq 1$ is a fixed integer,

$$\left\|\mathbb{E}\left(z_d^t\right)\right\| \;\leq\; c\,\frac{1\,-\,\rho^d}{d\,(1\,-\,\rho)\,\rho^{d-1}}\,\rho^t\,\|\mathbb{E}\left(\psi_0\right)\|. \tag{15b}$$

When $d = t$, the expected error converges sub-linearly with rate $1/t$,

$$\left\|\mathbb{E}\left(z_t^t\right)\right\| \;\leq\; c\,\frac{\rho\,(1-\rho^t)}{(1-\rho)\,t}\,\|\mathbb{E}(\psi^0)\| \tag{15c}$$

where c in (15) is the same positive constant.

While the bound on $\|\mathbb{E}\left(z_t^t\right)\|$ in (15c) is easily recovered from the bound on $\|\mathbb{E}\left(z_d^t\right)\|$ by setting $d = t$ in (15b), we observe significantly different convergence behavior. Namely, in contrast to $\|\mathbb{E}\left(x^t\right)\|$ and $\|\mathbb{E}\left(z_d^t\right)\|$, $\|\mathbb{E}\left(z_t^t\right)\|$ does not enjoy a linear convergence rate; rather, it converges to zero at a sub-linear rate $1/t$. While expected values of the algorithmic iterates and of the averaged output over the fixed window length d converge at the same linear rate ρ, the bound on $\|\mathbb{E}\left(z_d^t\right)\|$ is larger than the bound on $\|\mathbb{E}\left(x^t\right)\|$ by a factor that depends on ρ and d, $(1-\rho^d)/(d(1-\rho)\rho^{d-1})$. This factor is a monotonically increasing function of d, thereby suggesting that averaging over a longer window d increases the expected error $\|\mathbb{E}\left(z_d^t\right)\|$ at a given iteration t. Furthermore, since the second derivative of this factor with respect to d is always positive, the relative increase in the error bound in (15b) grows with additional increase in d.

Theorem 2 quantifies the effect of a window length on the steady-state variance of the averaged output z_d^t given by (4).

Theorem 2: Let the heavy-ball algorithm (2) with the strongly convex quadratic objective function f converge linearly with rate ρ. Then the variance $V_d^t := \mathbb{E}\left[\|z_d^t\|^2\right]$ converges to the steady-state value V_d^∞ as t goes to infinity. For a fixed window length $d \geq 1$, V_d^∞ is bounded by

$$\hat{V}_d^\infty(m) \,+\, \hat{V}_d^\infty(L) \;\leq\; V_d^\infty \;\leq\; (n-1)\hat{V}_d^\infty(m) \,+\, \hat{V}_d^\infty(L)$$

where n is the problem dimension, $x^t \in \mathbb{R}^n$, and $\hat{V}_d^\infty(m)$ and $\hat{V}_d^\infty(L)$ are the modal contributions to the steady-state variance associated with the smallest and the largest eigenvalues m and L of Q. In particular,

$$\hat{V}_d^\infty(m) \;=\; \frac{(1+\rho)^2 + 2\rho^{d+1}}{d\,(1-\rho)^4(1+\rho)^2} - \frac{4\rho(1+\rho+\rho^2)(1-\rho^d)}{d^2(1-\rho)^5(1+\rho)^3}$$

and $\hat{V}_d^\infty(L)$ is obtained by substituting $-\rho$ for ρ in the expression for $\hat{V}_d^\infty(m)$.

The bounds on V_d^∞ in Theorem 2 follow from $V_d^\infty = \sum_i \hat{V}_d^\infty(\lambda_i)$ and $\hat{V}_d^\infty(m) = \max_{\lambda_i} \hat{V}_d^\infty(\lambda_i)$, where $\hat{V}_d^\infty(\lambda_i)$ is the modal contribution to the steady-state variance of the ith eigenvalue of the matrix Q. When no averaging is done, i.e., for $d = 1$, our result matches the bounds in [18] with,

$$\hat{V}_1^\infty(m) \;=\; \hat{V}_1^\infty(L) \;=\; \frac{1\,+\,\rho^2}{(1\,-\,\rho^2)^3}.$$

We note that ρ given by (11a) determines the best achievable convergence rate of the heavy-ball algorithm in terms of κ. For this ρ, the product of $\hat{V}_1^\infty(m)$ and the settling time $T_s := 1/(1-\rho)$ has the following dependance on κ [18], [19],

$$\frac{\hat{V}_1^\infty(L)}{1\,-\,\rho} \;=\; \frac{\hat{V}_1^\infty(m)}{1\,-\,\rho} \;=\; \frac{(1\,+\,\sqrt{\kappa})^5\,(1\,+\,\kappa)}{64\kappa^{3/2}} \;=\; \Theta(\kappa^2).$$

We now show that, even for fixed $d > 1$, the product of the steady-state variance of the averaged output z_d^t, $V_d^\infty := \lim_{t\to\infty} \mathbb{E}\left[\|z_d^t\|^2\right]$, and the settling time scales as κ^2.

Corollary 1: Under the setting of Theorem 2, the product between the steady-state variance V_d^∞ and the settling time T_s scales as the square of the condition number κ,

$$\frac{V_d^\infty}{1\,-\,\rho} \;=\; \Theta(\kappa^2).$$

Since the largest contribution to the steady-state variance comes from the smallest eigenvalue m of the Hessian matrix Q, it is also of interest to examine the effect of window length d on $\hat{V}_d^\infty(m)$ for large values of the condition number κ. In

particular, for any fixed $d \geq 1$, as κ goes to infinity we have,

$$\lim_{\kappa\to\infty} \frac{1}{\kappa^2} \frac{\hat{V}_d^\infty(m)}{1-\rho} = \frac{1}{64}$$
$$\lim_{\kappa\to\infty} \frac{1}{\kappa} \left(\frac{\hat{V}_1^\infty(m)}{1-\rho} - \frac{\hat{V}_d^\infty(m)}{1-\rho} \right) = \frac{d^2-1}{192}.$$

Thus, for any fixed $d \geq 1$, $\hat{V}_d^\infty(m)/(1-\rho)$ is proportional to κ^2 regardless of the window length d. On the other hand, since $(\hat{V}_1^\infty(m) - \hat{V}_d^\infty(m))/(1-\rho)$ scales as $\kappa\, d^2$, for large values of the condition number averaging over longer window d leads to larger reduction in variance amplification relative to the standard heavy-ball algorithm.

Theorem 3 uncovers the influence of averaging over the entire algorithmic history on the variance of the averaged output z_t^t given by (4b).

Theorem 3: Let the heavy-ball algorithm (2) with the strongly convex quadratic objective function f converge linearly with rate ρ. Then the variance $V_t^t := \mathbb{E}\left[\|z_t^t\|^2\right]$ of the averaged output over the entire algorithmic history converges to zero at the sub-linear rate $1/t$,

$$\hat{V}_t^t(m) = \frac{1}{(1-\rho)^4\, t} - o(1/t)$$
$$\hat{V}_t^t(L) = \frac{1}{(1+\rho)^4\, t} + o(1/t).$$

Here, $\hat{V}_t^t(m)$ and $\hat{V}_t^t(L)$ are the modal contributions to the steady-state variance arising from the eigenvalues m and L of Q, respectively, and the higher order terms collected in $o(1/t)$ are always positive.

While the variance V_d^t grows monotonically in time, the variance V_t^t exhibits transient growth before asymptotic convergence to zero. This is due to the higher order terms collected in $o(1/t)$. However, even in the non-asymptotic regime, the variance associated with z_t^t is always smaller than the variance of z_d^t.

Lemma 2: Under the setting of Theorems 2 and 3, the variance V_t^d of the average over the entire algorithmic history z_t^t is smaller than or equal to the variance V_d^t of the moving window average z_d^t over the horizon of the fixed length d,

$$V_t^t \leq V_d^t, \quad \forall\, t \geq 1.$$

Compared to the averaging over a fixed window length d, Lemma 2 shows that, even in the non-asymptotic regime, averaging over the entire algorithmic history reduces variance.

Theorems 1 and 3 show that averaging over the entire algorithmic history eliminates steady-state variance while slowing convergence to a sub-linear rate. In particular, both the expected error and the variance of z_t^t approach zero at the rate $1/t$. On the other hand, a moving average over a fixed t-independent window of length d converges linearly in expectation while leading to a non-zero steady-state variance. This steady-state variance has the same scaling with the condition number as the variance of the output of the standard heavy-ball algorithm and the length of the averaging window d directly trades-off convergence speed with noise amplification. The moving average reduces the steady-state variance of z_d^t by a factor of approximately $1/d$ while increasing the t-independent factor in the bound on $\|\mathbb{E}\,(z_d^t)\,\|$ at a given iteration t.

IV. An example

We next provide an example to illustrate the merits and the effectiveness of averaging for the heavy-ball algorithm with parameters that optimize the convergence rate; see (11). For $n = 2$ with the diagonal Hessian matrix Q

$$Q = \begin{bmatrix} L & 0 \\ 0 & m \end{bmatrix}$$

the matrices $\hat{A}(\lambda_i)$ in (8) are given by

$$\hat{A}(m) = \begin{bmatrix} 0 & 1 \\ -\rho^2 & 2\rho \end{bmatrix}, \quad \hat{A}(L) = \begin{bmatrix} 0 & 1 \\ -\rho^2 & -2\rho \end{bmatrix}.$$

We set $\kappa = 10000$ which results in $\rho \approx 0.98$, and choose initial conditions $\hat{\psi}_1^0 = [1\ 1]^T$, $\hat{\psi}_2^0 = [1\ -1]^T$, which produce the same behavior in the expected error of $\hat{x}_i^t$ for $\lambda_1 = L$ and $\lambda_2 = m$.

Figure 1 illustrates the effect of different averaging schemes on the expected error and the modal contribution to the variance for $\lambda_1 = L$ and $\lambda_2 = m$. While in the absence of averaging (i.e., for $d = 1$) both the expected error and the variance behave identically for $\lambda_1 = L$ and $\lambda_2 = m$, averaging differently impacts outputs of these two subsystems. Although $\hat{A}(L)$ and $\hat{A}(m)$ both have repeated eigenvalues of magnitude ρ, the negative eigenvalues of $\hat{A}(L)$ cause $\hat{x}_1^t$ to change sign in every iteration. For this reason, averaging $\hat{x}_1^t$ of the subsystem with $\lambda_1 = L$ significantly reduces both the expected error and the variance.

Figure 2 highlights this effect. The figure on the right shows that even a short averaging window drastically reduces the steady-state variance $V_d^\infty(L)$. When d is strictly even or strictly odd, $V_d^\infty(L)$ is monotonically decreasing. The figure on the left illustrates that $\hat{V}_d^\infty(m)$ is a monotonically decreasing function of d. For large d the slope appears to be decreasing, indicating that the benefits of additional averaging drop off with increase in the window length d.

The advantages of even a small amount of averaging can further be seen in Figure 3, which compares the expected error and variance for $d = 10$ and $d = 100$. The figure on the right suggests that the improvement in the steady-state variance for window length $d = 10$ compared to $d = 1$ is more significant than the relative improvement of increasing window length further to $d = 100$. The figure on the left demonstrates that averaging over the entire algorithmic history results in significantly slower sub-linear convergence.

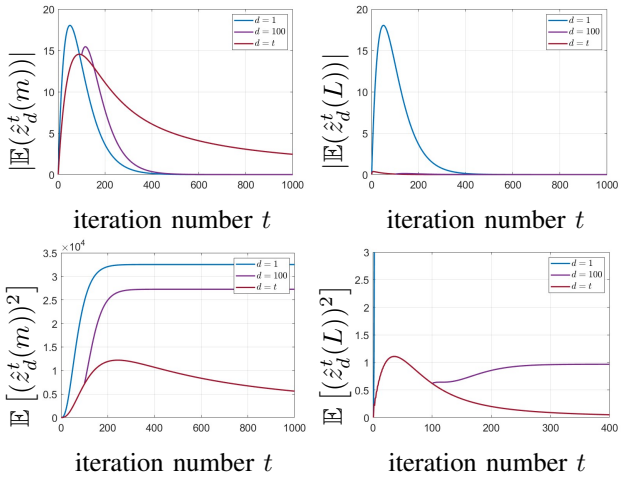

Fig. 1: Expected value and variance of the averaged output $\hat{z}^t(\lambda_i)$ at iteration t for subsystems associated with the eigenvalues m and L. We illustrate the effect of no averaging ($d = 1$), averaging over a moving window of length $d = 100$, and averaging over the entire history ($d = t$).

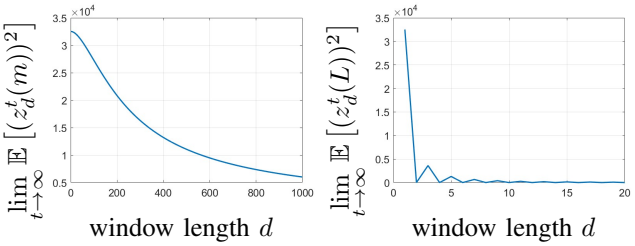

Fig. 2: Steady-state variance of the averaged output z_d^t, for subsystems associated with the eigenvalues m and L, as a function of the length of averaging window d. In both cases the variance approaches zero as d approaches infinity.

V. Concluding Remarks

We examine the impact of averaging on convergence properties and variance amplification of the heavy-ball algorithm for strongly convex quadratic problems. We show that averaging the algorithmic output over the entire algorithmic history results in zero steady-state variance at the expense of yielding a sub-linear convergence rate. On the other hand, averaging over a moving window of fixed length d reduces steady-state variance relative to the case without averaging while maintaining linear convergence rate. We demonstrate that the steady-state variance of the windowed average z_d^t is reduced by a factor of roughly $1/d$ compared to the steady-state variance of the non-averaged output x^t but the quadratic dependence on the condition number still remains. In our future work, we will extend our results to the class of two-step momentum algorithms and examine the effect of windowed averaging on transient behavior.

Appendix

A. Proof of Theorem 1

Proof: We establish rate of convergence results using geometric sums. For a fixed window length d, the error vector

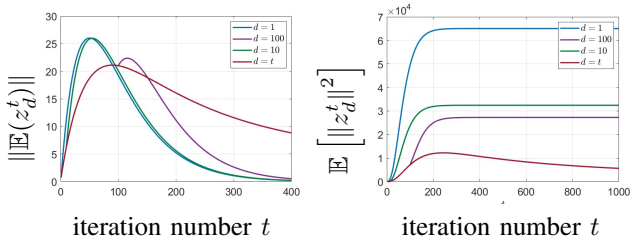

Fig. 3: Expected value and variance of the averaged output z_d^t for different values of the window length d and initial conditions $\hat{\psi}_1^0 = \hat{\psi}_2^0 = [1\ -1]^T$.

(z_d^t) is the linear combination of terms $x^t, \cdots, x^{t-d+1}$, the norms of which all approach zero at rate ρ. In particular, given the convergence bound introduced in (10), we can write

$$\begin{aligned}\left\|\mathbb{E}\left(z_d^t\right)\right\| \;&\leq\; \frac{1}{d}\sum_{k=t-d+1}^{t} c\,\rho^k\,\left\|\mathbb{E}(\psi^0)\right\| \\ &=\; c\,\rho^t\left(\frac{\rho^{1-d}\,(1-\rho^d)}{d\,(1-\rho)}\right)\left\|\mathbb{E}(\psi^0)\right\|\end{aligned}$$

where the additional constant factor is given by the partial geometric sum of ρ^k.

By the same technique we recover equation (15c), by considering the geometric sum of ρ^k from 1 to t.

■

B. Proof of Theorem 2

Proof: We first prove the results concerning the variance of the average over a fixed window length, z_d^t.

The modal contribution to variance of the averaged output at time t associated with eigenvalue λ_i of Q is given by

$$\hat{V}_d^t(\lambda_i) \;=\; \frac{1}{d^2}\sum_{k=t-d+1}^{t}\;\sum_{j=t-d+1}^{t}\mathbb{E}(\hat{x}_i^k\,\hat{x}_i^j) \qquad (16)$$

The covariance between outputs at times t, $t+j$ is given by

$$\mathbb{E}[\hat{x}_i^t\hat{x}_i^{t+j}] \;=\; \hat{C}\hat{P}_i^t(\hat{A}^j(\lambda_i))^T\hat{C}^T \qquad (17)$$

where $\hat{P}_i^t$ is the covariance matrix associated with the state $\hat{\psi}_i$ at time t and can be computed by

$$\hat{P}_i^t \;=\; \hat{P}_i^\infty \;-\; \hat{A}^t(\lambda_i)\,\hat{P}_i^\infty\,(\hat{A}^t(\lambda_i))^T. \qquad (18)$$

Based on the definition of $\hat{A}(\lambda_i)$ given in (8b), its eigenvalues μ_1 and μ_2 can be given as funtions of λ_i. We can use Lemma 1 of [10] to express $\hat{A}^t(\lambda_i)$ in terms of the eigenvalues μ_1, and μ_2 . Using Theorem 1 of [26] we can additionally write $\hat{P}_i^t$ in terms of eigenavlues μ_1, μ_2 .

We use expressions for $\hat{A}^t(\lambda_i)$ and $\hat{P}_i^t$ in terms of μ_1 and μ_2 and (18) to determine $\hat{P}_i^\infty$, which in combination with equations (16) and (17) can be used to express $\hat{V}_d^t(\lambda_i)$ as a function of window length d, time t and eigenvalues μ_1 and μ_2, which are in turn functions of the eigenvalues λ_i of the hessian Q. Taking the limit as $t \to \infty$ yields the expression for the steady-state variance $\hat{V}_d^\infty(\lambda_i)$. Expressions for $\hat{P}_i^t$, $\hat{V}_d^t(\lambda_i)$, and $\hat{V}_d^\infty(\lambda_i)$ are omitted due to length.

At $\lambda_1 = m$ and $\lambda_n = L$, we have $\mu_1 = \mu_2 = \rho$ and $\mu_1 = \mu_2 = -\rho$ respectively, and the above function simplifies to expression given in Theorem 2.

We will now show that $\hat{V}_d^\infty(m) \geq \hat{V}_d^\infty(\lambda_i)$ for $i = 1, \ldots, n$. For the heavy-ball algorithm, eigenvalues $\mu_{i,1}$ and $\mu_{i,1}$ are complex conjugates . It is straightforward to verify that for any complex conjugate eigenvalues μ_1, μ_2, the variance $\hat{V}_d^\infty(\lambda_i)$ is largest when $\mathrm{Im}(\mu_1) = \mathrm{Im}(\mu_2) = 0$.

By substituting μ for ρ in the expression $\hat{V}_d^\infty(m)$ given in Theorem 2, we obtain an expression for steady state variance as a function of a repeated eigenvalue μ. Using this expression it is straightforward to verify that for any real repeated eigenvalues $\mu_1 = \mu_2$, with $|\mu_{1,2}| \leq \rho$ the steady-state variance is maximized at $\mu_1 = \mu_2 = \rho$.

Thus $\hat{V}_d^\infty(m)$ is always larger than $\hat{V}_d^\infty(\lambda_i)$, which together with the definition

$$V_d^t \;=\; \sum_{i=1}^{n} \hat{V}_d^t(\lambda_i)$$

completes the proof. ∎

The corollary is easily verified by substituting the expression for ρ given in (11a) into the equation for $\hat{V}_d^\infty(m)$ given in Theorem 2 and computing the limit as κ goes to infinity. We determine

$$\frac{\hat{V}_d^\infty(m)}{1-\rho} \;=\; \frac{(1+\sqrt{\kappa})^5\, g(\kappa)}{128\, d^2\, \kappa^{3/2}}$$

where $g(\kappa)$ satisfies $\lim_{\kappa\to\infty} \frac{g(\kappa)}{\kappa} \;=\; 2\, d^2$.

C. Proof of Theorem 3

Proof: As in the previous proof, it is possible to express the modal contribution to variance $\hat{V}_t^t(\lambda_i)$ at time t as a function of the eigenvalues μ_1 and μ_2, which depend on λ_i and time t. We omit the expression due to excessive complexity. At $\lambda = m$ with $\mu_1 = \mu_2 = \rho$, $\hat{V}_t^t(m)$ can be expressed only in terms of ρ and t. Using the result from the previous proof with $d = t$, we confirm that for heavy ball parameters, $\hat{V}_t^t(m)$ is larger than $\hat{V}_t^t(\lambda_i)$ for all i. ∎

D. Proof of Lemma 2

Proof: We first show the result holds for every modal contribution to variance, so that $\hat{V}_t^t(\lambda_i) \leq \hat{V}_d^t(\lambda_i)$ for $i = 1, \cdots, n$. For any t such that $d \geq t$, we have $\hat{V}_t^t(\lambda_i) = \hat{V}_d^t(\lambda_i)$ by definition and the inequality is satisfied.

$\hat{V}_d^t(\lambda_i)$ is given in (16), with $\hat{V}_t^t(\lambda_i)$ given by the same expression evaluated at $d = t$. Based on (17) we conclude

$$\mathbb{E}(\hat{x}_i^{k_1}\hat{x}_i^j) \;\leq\; \mathbb{E}(\hat{x}_i^{k_2}\hat{x}_i^j)$$

for any $j \geq k_2 \geq k_1$. Thus any covariance term included in $\hat{V}_t^t(\lambda_i)$ but not included in $\hat{V}_d^t(\lambda_i)$, i.e. any $\mathbb{E}(\hat{x}_i^k\, \hat{x}_i^j)$ such that either j or k is less than $t - d + 1$, must be less than or equal to a term included in $\hat{V}_d^t(\lambda_i)$. Thus we have $\sum_{i=1}^n \hat{V}_d^t(\lambda_i) \;\geq\; \sum_{i=1}^n \hat{V}_t^t(\lambda_i)$ and completes the proof.

∎

REFERENCES

[1] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*. Springer Science & Business Media, 2013, vol. 87.

[2] Y. Nesterov, "Gradient methods for minimizing composite objective functions," *Math. Program.*, vol. 140, no. 1, pp. 125–161, 2013.

[3] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*. Kluwer Academic Publishers, 2004, vol. 87.

[4] B. T. Polyak, "Some methods of speeding up the convergence of iteration methods," *USSR Comput. Math. & Math. Phys.*, vol. 4, no. 5, pp. 1–17, 1964.

[5] Y. Nesterov, *Lectures on convex optimization*. Springer Optimization and Its Applications, 2018, vol. 137.

[6] L. Lessard, B. Recht, and A. Packard, "Analysis and design of optimization algorithms via integral quadratic constraints," *SIAM J. Optim.*, vol. 26, no. 1, pp. 57–95, 2016.

[7] B. V. Scoy, R. A. Freeman, and K. M. Lynch, "The fastest known globally convergent first-order method for minimizing strongly convex functions," *IEEE Control Syst. Lett.*, vol. 2, no. 1, pp. 49–54, 2018.

[8] A. Badithela and P. Seiler, "Analysis of the heavy-ball algorithm using integral quadratic constraints," in *Proceedings of the 2019 American Control Conference*. IEEE, 2019, pp. 4081–4085.

[9] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *Proc. ICML*, 2013, pp. 1139–1147.

[10] H. Mohammadi, S. Samuelson, and M. R. Jovanović, "Transient growth of accelerated optimization algorithms," *IEEE Trans. Automat. Control*, vol. 68, no. 3, pp. 1823–1830, 2023.

[11] Y. Bengio, "Gradient-based optimization of hyperparameters," *Neural Computation*, vol. 12, pp. 1889–1900, 2000. [Online]. Available: https://api.semanticscholar.org/CorpusID:5671899

[12] A. Beirami, M. Razaviyayn, S. Shahrampour, and V. Tarokh, "On optimal generalizability in parametric learning," in *NIPS*, 2017.

[13] Z.-Q. Luo and P. Tseng, "Error bounds and convergence analysis of feasible descent methods: a general approach," *Ann. Oper. Res.*, vol. 46, no. 1, pp. 157–178, 1993.

[14] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Statist.*, pp. 400–407, 1951.

[15] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, "Robust stochastic approximation approach to stochastic programming," *SIAM J. Optim.*, vol. 19, no. 4, pp. 1574–1609, 2009.

[16] O. Devolder, "Exactness, inexactness and stochasticity in first-order methods for large-scale convex optimization," Ph.D. dissertation, Louvain-la-Neuve, 2013.

[17] P. Dvurechensky and A. Gasnikov, "Stochastic intermediate gradient method for convex problems with stochastic inexact oracle," *J. Optimiz. Theory App.*, vol. 171, no. 1, pp. 121–145, 2016.

[18] H. Mohammadi, M. Razaviyayn, and M. R. Jovanović, "Robustness of accelerated first-order algorithms for strongly convex optimization problems," *IEEE Trans. Automat. Control*, vol. 66, no. 6, pp. 2480–2495, June 2021.

[19] H. Mohammadi, M. Razaviyayn, and M. R. Jovanović, "Tradeoffs between convergence rate and noise amplification for momentum-based accelerated optimization algorithms," 2022, submitted; also arXiv:2209.11920.

[20] B. V. Scoy and L. Lessard, "The speed-robustness trade-off for first-order methods with additive gradient noise," 2021, arXiv:2109.05059.

[21] B. Polyak, "Comparison of convergence rate of one-step and multistep optimization algorithms in the presence of noise," *Izv. Akad. Nauk SSSR, Tekh. Kibern.*, vol. 1, pp. 9–12, 01 1977.

[22] B. Polyak and A. Juditsky, "Acceleration of stochastic approximation by averaging," *SIAM Journal on Control and Optimization*, vol. 30, pp. 838–855, 07 1992.

[23] S. Gadat, F. Panloup, and S. Saadane, "Stochastic heavy ball," *Electronic Journal of Statistics*, vol. 12, no. 1, pp. 461 – 529, 2018. [Online]. Available: https://doi.org/10.1214/18-EJS1395

[24] P. Jain, S. M. Kakade, R. Kidambi, P. Netrapalli, and A. Sidford, "Accelerating stochastic gradient descent for least squares regression," in *Conference On Learning Theory*. PMLR, 2018, pp. 545–604.

[25] M. Danilova and G. Malinovskiy, "Averaged heavy-ball method," *Computer Research and Modeling*, vol. 14, pp. 277–308, 04 2022.

[26] H. Mohammadi, M. Razaviyayn, and M. R. Jovanović, "Variance amplification of accelerated first-order algorithms for strongly convex quadratic optimization problems," in *Proceedings of the 57th IEEE Conference on Decision and Control*, Miami, FL, 2018, pp. 5753–5758.