2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Performance of noisy three-step accelerated first-order optimization algorithms for strongly convex quadratic problems

Samantha Samuelson, Hesameddin Mohammadi, and Mihailo R. Jovanović

Abstract— **We study the class of first-order algorithms in which the optimization variable is updated using information from three previous iterations. While two-step momentum algorithms akin to heavy-ball and Nesterov's accelerated methods achieve the optimal convergence rate, it is an open question if the three-step momentum method can offer advantages for problems in which exact gradients are not available. For strongly convex quadratic problems, we identify algorithmic parameters which achieve the optimal convergence rate and examine how additional momentum terms affects the trade-offs between acceleration and noise amplification. Our results suggest that for parameters that optimize the convergence rate, introducing additional momentum terms does not provide improvement in variance amplification relative to standard accelerated algorithms.**

Index Terms— **Convex optimization, gradient descent, heavy-ball method, Nesterov's accelerated algorithms, noisy gradients, performance tradeoffs.**

I. Introduction

Accelerated first-order optimization algorithms [1]–[4] are widely used in a variety of large-scale optimization settings [5]–[7], due to their favorable asymptotic behavior [8]–[12] while maintaining low per-iteration complexity. The trade-off between acceleration and robustness has been well studied [13]–[20], determining that increased acceleration comes at the price of increased sensitivity to noise. Previous work [21], [22] establishes a fundamental limitation on the product of noise amplification and settling time imposed by condition number, and [23] examines a parameterized family of two-step momentum algorithms that enable systematic trade-offs between these quantities.

We extend [22] by analyzing the steady-state variance of the error in the optimization variable in the presence of additive white noise perturbing the iterations for a class of three-step accelerated algorithms, where the current estimate of the optimal solution x^t is updated using three previous iterates. This type of noise is used to model uncertainty due to roundoff, quantization, and communication errors [24]. For strongly convex quadratic problems, we analyze the effect of the additional momentum term on sensitivity to noise by providing upper and lower bounds on variance amplification in terms of the convergence rate.

Our results show that additional momentum stretches the distance between upper and lower bounds on noise amplification. Among the family of parameters which achieve the optimal rate of convergence, the smallest worst-case noise amplification corresponds to the standard heavy-ball algorithm, thus an additional momentum term is not advantageous. Adding additional momentum offers no advantage in terms of convergence rate, and increases the maximal achieved contribution to noise amplification. While the minimal noise amplification does decrease slightly, the scale is insignificant compared to the increase in the worst-case. Ultimately we conclude that while adding an additional history term to the standard gradient decent scheme can be beneficial, including further history terms is offers no advantage in convergence rate or steady state variance.

The rest of the paper is structured as follows. In Section II, we provide the problem formulation. In Section III, we present our results regarding convergence rate and steady-state variance amplification. In Section IV we provide conditions for stability and linear convergence along with the proofs of all results.

II. Motivation and background

We study convergence rate and noise amplification of first-order algorithms for unconstrained strongly convex optimization problems

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

In particular, we examine the class of algorithms that utilize information from three previous iterations to update the optimization variable x^t,

$$\begin{aligned} x^{t+3} \;=\; & \beta_2 x^{t+2} \,+\, \beta_1 x^{t+1} \,+\, \beta_0 x^t \,-\, \\ & \alpha \nabla f\big(\gamma_2 x^{t+2} + \gamma_1 x^{t+1} + \gamma_0 x^t\big) \,+\, w^t \end{aligned} \tag{2}$$

Here t is the iteration index, α is the stepsize, β_k and γ_k are the algorithmic parameters, and w^t is a white noise with

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

First-order optimality conditions impose the following constraints on parameters β_k and γ_k

$$\sum_{k\,=\,0}^{2} \beta_k \,=\, 1, \quad \sum_{k\,=\,0}^{2} \gamma_k \,=\, 1. \tag{4}$$

Under these conditions, for $\gamma_0 = \beta_0 = 0$, we recover familiar first-order algorithms with the following choices of remaining parameters: (i) gradient descent ($\gamma_1 = \beta_1 = 0$, $\gamma_2 = \beta_2 = 1$); (ii) Polyak's heavy-ball method ($\gamma_1 = 0$, $\gamma_2 = 1$); and (iii) Nesterov's accelerated algorithm ($\gamma_1 = \beta_1$, $\gamma_2 = \beta_2$).

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

979-8-3503-0123-6/23/$31.00 ©2023 IEEE

Relative to these standard algorithms, we introduce additional momentum terms in (2) to examine tradeoffs between convergence rate and noise amplification. While two-step momentum algorithms achieve the optimal convergence rate [8], it is an open question if the three-step momentum method can offer advantages in terms of noise amplification.

In this paper, we study this question 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.

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,

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

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

$$\psi^t \,=\, \left[\begin{array}{ccc} (y^t)^T & (y^{t+1})^T & (y^{t+2})^T \end{array}\right]^T \tag{6b}$$

and A, B, C are constant matrices determined by,

$$\begin{array}{rcl} A &=& \left[\begin{array}{ccc} 0 & I & 0 \\ 0 & 0 & I \\ -D_0 & -D_1 & -D_2 \end{array}\right] \\ B &=& \left[\begin{array}{ccc} 0 & 0 & I \end{array}\right]^T,\; C \,=\, \left[\begin{array}{ccc} I & 0 & 0 \end{array}\right] \\ D_k &=& \alpha\gamma_k Q \,-\, \beta_k I,\;\; k \,=\, \{0,1,2\}. \end{array} \tag{6c}$$

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 $Q \in \mathbb{R}^{n\times n}$,

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

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

$$\begin{array}{rcl} \hat{A}(\lambda_i) &=& \left[\begin{array}{ccc} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -d_0(\lambda_i) & -d_1(\lambda_i) & -d_2(\lambda_i) \end{array}\right] \\ \hat{B} &=& \left[\begin{array}{ccc} 0 & 0 & 1 \end{array}\right]^T,\; \hat{C} \,=\, \left[\begin{array}{ccc} 1 & 0 & 0 \end{array}\right] \\ d_k(\lambda_i) &=& \alpha\gamma_k\lambda_i \,-\, \beta_k,\;\; k \,=\, \{0,1,2\}. \end{array} \tag{8b}$$

Since the two-step momentum algorithm is obtained by setting $d_0(\lambda_i) = 0$ for each i, it is of interest to examine the influence of $d_0(\lambda_i)$ on the convergence rate and noise amplification.

B. Convergence rate and noise amplification

System (6) is stable if eigenvalues of matrices $\hat{A}(\lambda_i)$ have real parts with absolute value less than one for each $i = 1,\ldots,n$. The characteristic polynomial of $\hat{A}(\lambda_i)$ in (8b) is determined by,

$$F_{\lambda_i}(z) \,=\, z^3 \,+\, d_2(\lambda_i)z^2 \,+\, d_1(\lambda_i)z \,+\, d_0(\lambda_i). \tag{9}$$

In Section IV, we utilize the Jury stability criterion [25] to determine the conditions for stability and ρ-exponential stability of system (6). We recall that exponential stability implies

$$\|\psi^t \,-\, \psi^\star\| \,\leq\, c\rho^t\|\psi^0 \,-\, \psi^\star\| \tag{10}$$

where the convergence rate is given by

$$\rho \,=\, \max_{\lambda\in[m,L]}\,|\mathrm{eig}(\hat{A}(\lambda))|. \tag{11}$$

For LTI system (6), the solution of the algebraic Lyapunov equation

$$P \,=\, APA^T \,+\, BB^T \tag{12}$$

can be used to compute the steady-state variance of the error in the optimization variable,

$$J \,:=\, \lim_{t\,\to\,\infty}\frac{1}{t}\sum_{k=0}^{t}\mathbb{E}\left(\|x^k \,-\, x^\star\|^2\right) \,=\, \sum_{i\,=\,1}^{n}\hat{J}(\lambda_i). \tag{13}$$

Here, P denotes the steady-state covariance matrix of ψ^t

$$P \,=\, \lim_{t\,\to\,\infty}\mathbb{E}\left[\psi^t(\psi^t)^T\right] \tag{14}$$

and an explicit expression for the contribution of the ith eigenvalue λ_i of Q to the variance amplification,

$$\hat{J}(\lambda) \,=\, \mathrm{trace}\left(\hat{C}\hat{P}(\lambda_i)\hat{C}^T\right). \tag{15}$$

is obtained in Section IV-A by solving the decoupled family of algebraic Lyapunov equations for $\hat{P}(\lambda_i)$.

III. Main Results

We next summarize our main results. In Theorem 1, we present general bounds on the smallest and largest modal contributions to the variance amplification

$$\hat{J}_{\min} \,:=\, \min_{\lambda}\hat{J}(\lambda),\;\;\; \hat{J}_{\max} \,:=\, \max_{\lambda}\hat{J}(\lambda)$$

for any set of stabilizing parameters

$$\theta \,:=\, \{\alpha,\beta_0,\beta_1,\beta_2,\gamma_0,\gamma_1,\gamma_2\}. \tag{16}$$

In Theorem 2, we characterize the set of parameters that achieve the optimal rate of convergence and, in Theorem 3, we provide bounds for $\hat{J}_{\min}$ and $\hat{J}_{\max}$ for these parameters.

We first present upper and lower bounds on modal contributions $\hat{J}(\lambda)$ to the variance amplification for any set of stabilizing parameters θ.

Theorem 1: Let the parameters θ be such that the three-step momentum algorithm (6) achieves the linear convergence rate ρ for all $f \in \mathcal{Q}_m^L$. Then the modal contribution $\hat{J}(\lambda)$ to the steady-state variance amplification satisfies

$$\frac{16\rho^2}{(1+\rho)^5} \leq \hat{J}(\lambda) \leq \frac{1+4\rho^2+\rho^4}{(1-\rho^2)^5}. \tag{17}$$

In contrast, for the two-step momentum algorithm we have

$$\frac{4\rho}{(1+\rho)^3} \leq \hat{J}(\lambda) \leq \frac{1+\rho^2}{(1-\rho^2)^3}. \tag{18}$$

Introducing the third momentum term d_0 decreases the lower bound while increasing the upper bound. Essentially, additional momentum widens the range of best-case and worst-case noise amplification, as we would expect to result from the introduction of an additional degree of freedom.

We now examine the effect of additional momentum terms on noise amplification for parameters designed to optimize convergence rate. First, we describe the parameters that achieve this rate for the three-step momentum algorithm.

Theorem 2: For strongly convex quadratic objective function $f \in \mathcal{Q}_m^L$ with the condition number $\kappa := L/m$, the optimal convergence rate of the three-step momentum algorithm (6) is given by

$$\rho \,=\, 1 \,-\, \frac{2}{\sqrt{\kappa}\,+\,1}.$$

This convergence rate is only achieved by the following set of parameters,

$$\begin{array}{lll} \beta_0 \,=\, -d_0, & \beta_1 \,=\, \frac{-\rho^4+d_0\rho^2+d_0}{\rho^2} & \beta_2 \,=\, \frac{\rho^4+\rho^2-d_0}{\rho^2} \\ \gamma_0 \,=\, 0 & \gamma_1 \,=\, \frac{d_0}{\rho^2+d_0} & \gamma_2 \,=\, \frac{\rho^2}{\rho^2+d_0} \\ \alpha \,=\, \frac{4\rho+4d_0\rho^{-1}}{L-m}, & d_0 \,\in\, [-\rho^3,\, \rho^3]. & \end{array} \tag{19}$$

The parameters in Theorem 2 are expressed in terms of d_0, where stability requirements impose $d_0 \in [-\rho^3, \rho^3]$; see Section IV-A. The optimal convergence rate matches the one achieved by the Polyak's heavy-ball method (which is recovered for $d_0 = 0$). Furthermore, our result implies that any set of stabilizing parameters θ with $\gamma_0 \neq 0$, which allows $d_0(\lambda)$ to vary with λ, yields slower rate of convergence.

We next present noise amplification bounds for the algorithmic parameters provided in Theorem 2.

Theorem 3: For the three-step momentum algorithm (6) with the parameters provided in Theorem 2, the modal contribution $\hat{J}(\lambda)$ to the steady-state variance amplification satisfies,

$$\frac{1+\rho+\rho^2}{2(1+\rho)^5} \leq \hat{J}(\lambda) \leq \frac{1+4\rho^2+\rho^4}{(1-\rho^2)^5}. \tag{20}$$

For the two-step momentum algorithm we have

$$\frac{1}{1-\rho^4} \leq \hat{J}(\lambda) \leq \frac{1+\rho^2}{(1-\rho^2)^3}. \tag{21}$$

Since the optimal convergence rate ρ in Theorem 2 depends on the condition number κ, the above bounds can be expressed in terms of κ. The following corollary extends the bounds on the product between modal contributions to the variance amplification and the settling time for the two-step momentum method [21], [22] to the three-step momentum method with parameters that optimize the convergence rate.

Corollary 1: For the three-step momentum algorithm (6) with the parameters provided in Theorem 2, the product between modal contribution to the variance amplification $\hat{J}(\lambda)$ and the settling time $T_s = 1/(1-\rho)$ satisfies

$$\mathcal{O}(\sqrt{\kappa}) \,\leq\, \hat{J}(\lambda) \times T_s \,\leq\, \mathcal{O}(\kappa^3). \tag{22}$$

For the two-step momentum method we have

$$\mathcal{O}(\kappa) \,\leq\, \hat{J}(\lambda) \,\times T_s \,\leq\, \mathcal{O}(\kappa^2). \tag{23}$$

Corollary 1 shows that the interval to which $\hat{J}(\lambda) \times T_s$ belongs widens for the three-step momentum algorithm. Since the upper bound in (22) is tight, for the parameters that optimize the convergence rate the variance amplification increases relative to the two-step momentum method.

The next section defines the ρ-convergence region and provides proofs of all results.

IV. Proofs

A. Defining the ρ-convergence region

We first describe the stability region as defined by parameters d_k defined in (8b). System (8) is stable when the roots of the characteristic equation (9) have absolute value less than one. The Jury stability criterion [25] applied to the characteristic polynomial (9) consist of the following necessary conditions

$$\begin{array}{l} F(1) \,=\, 1 \,+\, d_2 \,+\, d_1 \,+\, d_0 \,>\, 0 \\ (-1)^3 F(-1) \,=\, 1 \,-\, d_2 \,+\, d_1 \,-\, d_0 \,>\, 0 \\ |d_0| \,<\, 1 \\ |d_0^2 \,-\, 1| \,>\, |d_0 d_2 \,-\, d_1| \end{array} \tag{24}$$

which motivate us to define the following values

$$\begin{array}{l} a := 1 - d_0 \,+\, d_1 \,-\, d_2 \,\geq\, 0 \\ b := 1 \,+\, d_0 \,+\, d_1 \,+\, d_2 \,\geq\, 0 \\ c := 1 \,-\, d_0^2 \,-\, d_1 \,+\, d_0 d_2 \,\geq\, 0 \\ d := 1 \,-\, d_0^2 \,\geq\, 0 \end{array} \tag{25}$$

which must be positive to guarantee stability. The resulting three dimensional stability region is shown in Figure 1a.

Repeating the process for the scaled characteristic equation

$$F_\lambda(\rho z) \,=\, \rho^3 z^3 \,+\, d_2(\lambda)\rho^2 z^2 \,+\, d_1(\lambda)\rho z \,+\, d_0(\lambda) \tag{26}$$

yields the conditions

$$\rho^3 \,-\, d_2\rho^2 \,+\, d_1\rho \,-\, d_0 \,\geq\, 0 \tag{27a}$$

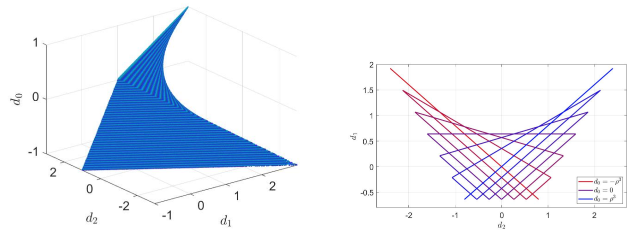
$$\rho^3 \,+\, d_2\rho^2 \,+\, d_1\rho \,+\, d_0 \,\geq\, 0 \tag{27b}$$

$$\rho^6 \,-\, d_1\rho^4 \,+\, d_0 d_2\rho^2 \,-\, d_0^2 \,\geq\, 0 \tag{27c}$$

$$\rho^6 \,-\, d_0^2 \,\geq\, 0 \tag{27d}$$

for ρ-convergence.

While the ρ-convergence region is non-convex, the non-convexity appears exclusively in d_0 as seen in Figure 1a. If



(a) (b)

Fig. 1: Figure (a) shows the three dimensional stability region in d_0, d_1, and d_2. Different shades of blue correspond to the level sets of d_0. Figure (b) shows the d_0 level sets $\Delta_\rho(d_0)$ of ρ-convergence region for $\rho = 0.8$, $d_0 \in [-\rho^3,\ \rho^3]$. At $d_0 = \pm\rho^3$, the convergence region collapses to a single line, and at $d_0 = 0$, $\Delta_\rho(0)$ recovers the two step ρ-convergence region.

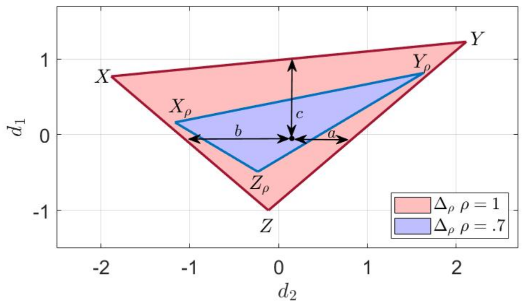


Fig. 2: Level sets of the stability region and the ρ-convergence region at $\rho = 0.7$, at $d_0 = 0.1$, as defined by the positivity constraints of (25) and (27).

we examine the stability region for a fixed d_0, the region is a convex triangle defined by first three constrains given in (27), denoted $\Delta_\rho(d_0)$, as seen in Figure 2. As expected, setting $d_0 = 0$ recovers the two dimensional stability triangle described in [22].

We have labeled the vertices of the ρ-convergence region X_ρ, Y_ρ, and Z_ρ as shown in Figure 2. For a fixed d_0, the d_1, d_2 coordinates of these vertices are given by

$$\begin{array}{lll} Z_\rho: & d_2 = -d_0\rho^{-2}, & d_1 = -\rho^2 \\ Y_\rho: & d_2 = 2\rho + d_0\rho^{-2}, & d_1 = \rho^2 + 2d_0\rho^{-1} \\ X_\rho: & d_2 = -2\rho + d_0\rho^{-2}, & d_1 = \rho^2 - 2d_0\rho^{-1}. \end{array} \tag{28}$$

Similarly, the edges of the stability and ρ-convergence regions occur along the lines where the constraints in (25) and (27) respectively are exactly zero.

From this we observe that the values a, b, and c as defined in (25) have intuitive geometric interpretations as seen in Figure 2. We can now use these values to express $\hat{J}(\lambda)$ in terms of a, b, and c.

Lemma 1: For strongly convex quadratic objective function $f \in \mathcal{Q}_m^L$, the modal contribution $\hat{J}(\lambda)$ to the steady-state variance amplification of system (8) with stabilizing parameters θ, is given by

$$\hat{J}(\lambda) = \frac{(1 + d_0(\lambda))a(\lambda) + (1 - d_0(\lambda))b(\lambda)}{2a(\lambda)b(\lambda)c(\lambda)} \tag{29}$$

The Lemma is proven by solving (12) for the decoupled system given in (8), and simplifying the result using the definitions of a, b, and c in (25).

B. *Proof of Theorem 1*

We first establish the upper bound on $\hat{J}(\lambda)$ given in both Theorems 1 and 3.

Proof: For a fixed d_0, $\hat{J}$ is convex in a, b, c on the positive orthant, which is required for stability. Given that a, b, c are affine functions of d_1 and d_2, $\hat{J}$ is convex in d_1, d_2 and must achieve it's maximum at one of the vertices X_ρ, Z_ρ, Y_ρ.

The (d_1, d_2) coordinates of the vertices defined in (28) allow us to determine $\hat{J}(\lambda)$ at each vertex. Exact function forms are omitted due to length. We define the following maximal value of $\hat{J}(\lambda)$ for fixed d_0

$$\hat{J}_{\max}(d_0) \;=\; \frac{\rho^4\left(2|d_0|\rho(1-\rho^2) + (\rho^2 - d_0^2)(1+\rho^2)\right)}{(\rho^4 - d_0^2)(\rho - |d_0|)^2(1-\rho^2)^3} \tag{30}$$

achieved by $\hat{J}$ at Y_ρ for $d_0 \geq 0$ and at X_ρ for $d_0 \leq 0$, which is in turn maximized at $d_0 = -\rho^3$

$$\max_{d_0}\ \max_{\lambda}\ \hat{J}(\lambda) \;=\; \frac{1 + 4\rho^2 + \rho^4}{(1-\rho^2)^5}. \tag{31}$$

■

We will now prove the lower bound given in Theorem 1. Consider the decomposition

$$\begin{aligned} \hat{J}(\lambda) \;&=\; F_{bc}(\lambda) \;+\; F_{ac}(\lambda) \\ F_{bc}(\lambda) \;&:=\; \frac{(1 + d_0(\lambda))}{2b(\lambda)c(\lambda)}, \quad F_{ac}(\lambda) \;:=\; \frac{(1 - d_0(\lambda))}{2a(\lambda)c(\lambda)} \end{aligned} \tag{32}$$

where in future references we drop the dependence on λ for ease of notation.

Proof: The lower bound is obtained by bounding $F_{ac}(\lambda)$ and $F_{bc}(\lambda)$, introduced in (32), independently, taking advantage of the inequality

$$\min_{\lambda}\left[F_{ac}(\lambda) + F_{bc}(\lambda)\right] \;\geq\; \min_{\lambda}\left[F_{ac}(\lambda)\right] + \min_{\lambda}\left[F_{bc}(\lambda)\right]. \tag{33}$$

For any (d_1, d_2) interior to $\Delta_\rho(d_0)$ it is possible to move along the line of constant c and increase either a or b until the either the $X_\rho Z_\rho$ or the $Z_\rho Y_\rho$ edge of $\Delta_\rho(d_0)$ is reached. Thus F_{ac} and F_{bc} must each be minimized along the $X_\rho Z_\rho$ and $Z_\rho Y_\rho$ edges of $\Delta_\rho(d_0)$ respectively. Using the equality constraints which define these edges, we present the following lower bounds on F_{ac} and F_{bc}.

Proposition 1: For any $d_0 \in [-\rho^3,\ \rho^3]$, F_{ac} and F_{bc} are lower bounded by

$$F_{ac} \geq \frac{8\rho^2}{(1+\rho)^5} \qquad F_{bc}(c) \geq \frac{8\rho^2}{(1+\rho)^5}. \tag{34}$$

Proof: Using the relations given in (27) it is possible to determine the minimums of F_{ac} and F_{bc} for a fixed d_0, and then further minimize the results with respect to d_0. ■

The propositions above combined with (33) complete the proof. By evaluating the minimums of F_{ac} along $X_\rho Z_\rho$ and F_{bc} along $Z_\rho Y_\rho$ at $d_0 = 0$, and evaluating the upper bound $\hat{J}_{\max}(d_0)$ given in (30) at $d_0 = 0$ provides upper and lower bounds for the two step case.

■

C. Proof of Theorem 2

Proof: Choosing parameters β_k and γ_k amounts to placing a line in 3-D space of d_0, d_1, d_2, parameterized by λ, which must remain within the ρ-convergence region defined in (27), as seen in Figure 1a. Based on the endpoints of this line at eigenvalues m and L, the largest permissible condition number κ is determined by

$$\kappa = \frac{\alpha L}{\alpha m} = \frac{1 + d_0(L) + d_1(L) + d_2(L)}{1 + d_0(m) + d_1(m) + d_2(m)}, \tag{35}$$

based on the equalities in (4). In order to maximize κ for a given ρ, we wish to choose endpoints which maximize this ratio.

We will now determine parameters which achieve optimal rate of convergence by considering endpoints $(d_0, d_1, d_2)(m)$ and $(d_0, d_1, d_2)(L)$ which achieve $\kappa = \frac{(1+\rho)^2}{(1-\rho)^2}$.

In the case $d_0(\lambda)$ is fixed, with $\gamma_0 = 0$, it is straightforward to verity the optimal line placement lies along the $X_\rho Y_\rho$ edge of the ρ-convergence region, defined (27c).

Using the (d_1, d_2) coordinates of endpoints X_ρ and Y_ρ given in (28) in conjunction with (35) yields $\kappa = \frac{(1+\rho)^2}{(1-\rho)^2}$ which is independent of d_0 and matches the optimal. By solving the equations $d_k(\lambda) = -\beta_k + \alpha\lambda\gamma_k$ at m and L we produce the parameters given in (19).

We will now examine the case where γ_0 is strictly non-zero. Suppose we are given $d_0(m)$ and $d_0(L)$.

Considering the definition of κ given in (35), it is evident that increasing $d_0(L)$ should increase κ; however as $d_0(L)$ increases the convergence region $\Delta_\rho(L)$ shifts, resulting in a decrease in $d_1(L)$ and $d_2(L)$. In order to quantify this trade-off, we consider the following question: As $d_0(L)$ is increased to $d_0(m) + \Delta_{d_0}$, how are $(d_1(L), d_2(L))$ affected?

Given that we wish $(d_0, d_1, d_2)(L)$ to be greater than $(d_0, d_1, d_2)(m)$, it suffices to consider lines with the m endpoint on the $X_\rho Z_\rho$ edge and the L endpoint along the $Z_\rho Y_\rho$ edge and $d_0(L) > d_0(m)$.

As we increase $\alpha\lambda$ incrementally by ϵ_λ, d_1 must continue to satisfy the constraint (27c) which requires

$$\begin{aligned} d_1(m + \epsilon\lambda) \; = \; d_1(m) + \gamma_1\epsilon_\lambda \; \leq & \\ \rho^2 + (d_0(m) + \gamma_0\epsilon_\lambda)(d_2(m) + \gamma_2\epsilon_\lambda)\rho^{-2} & \\ - \, (d_0(m) + \gamma_0\epsilon_\lambda)^2\rho^{-4} & \end{aligned} \tag{36}$$

and results in following the constraint on γ_1 in terms of γ_0

$$\begin{aligned} \gamma_1 \; \leq \; & [d_1^{\max}(m) - d_1(m)] \\ & + d_0(m)\gamma_2\rho^{-2} + d_2(m)\gamma_0\rho^{-2} - 2d_0\gamma_0\rho^{-4}. \end{aligned} \tag{37}$$

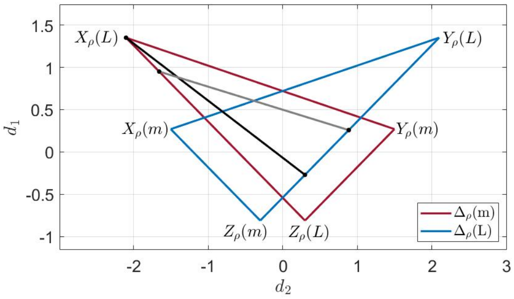


Fig. 3: Overlaid d_0 level sets of $\Delta_\rho(d_0)$ at $\rho = 0.9$, for $d_0(m) = -\rho^3/3$ and $d_0(L) = \rho^3/3$ in red and blue respectively. In black and gray we see two examples of a parameterized line $(d_2(\lambda), d_1(\lambda))$ which runs from the $X_\rho(m)Z_\rho(m)$ edge to the $Z_\rho(L)Y_\rho(L)$ edge. We can see that in order to satisfy the constraint (27c) as d_0 changes, the d_1/d_2 slope must be more negative than one might expect, and $(d_1, d_2)(L)$ are both smaller than they would be at the $Y_\rho(m)$ vertex.

Using the relations

$$\begin{array}{lcl} \gamma_2 & = & 1 - \gamma_1 - \gamma_0 \\ d_1^{\max} & = & \rho^2 - 2d_0(m)\rho^{-1} \\ d_2(m) & = & -\rho - d_1(m)\rho^{-1} - d_0(m)\rho^{-2} \end{array} \tag{38}$$

we can express (37) as a functions solely of $d_0(m)$, $d_1(m)$, and ρ.

Setting $(d_1, d_2)(m)$ along $X_\rho Z_\rho$ and $(d_1, d_2)(L)$ along $Z_\rho Y_\rho$, together with definitions of γ_k, requires

$$\begin{aligned} & 2\rho^3 + 2d_1(m)\rho - \Delta_{d_2}\rho^2 + \Delta_{d_1}\rho - \Delta_{d_0} \; = \; 0 \\ & \gamma_0\Delta_{d_1} = \gamma_1\Delta_{d_0} \\ & \gamma_1\Delta_{d_2} = (1 - \gamma_1 - \gamma_0)\Delta_{d_1}. \end{aligned} \tag{39}$$

We can now solve the system of equations given in (37) and (39), where we have chosen to set γ_1 equal to its upper bound, as in order to maximize $d_1(L)$ an $d_2(L)$ we wish to make $\gamma_1/(1 - \gamma_1 - \gamma_0)$ large, which is achieved by making γ_1 as large as allowed.

Thus we obtain γ_0, γ_1, Δ_{d_1} and Δ_{d_2} as functions of $d_0(m)$, $d_1(m)$ and Δ_{d_0}. The definitions are not included due to complexity. Along the $X_\rho Y_\rho$ edge, with $\Delta_{d_0} = 0$, we have $\Delta_{d_1} = 4d_0\rho^{-1}$ and $\Delta_{d_2} = 4\rho$. Using the symbolic computation engine Mathematica we can verify that, for $d_1(m) \in [-\rho^2, \rho^2 - d_0(m)\rho^{-1}]$,

$$(\Delta_{d_0} + \Delta_{d_1} + \Delta_{d_2}) \; \leq \; 4\rho + 4d_0(m)\rho^{-1}$$

with equality only in the case $\Delta_{d_0} = 0$ and $d_1(m) = \rho^2 - d_0(m)\rho^{-1}$, the d_1 coordinate of X_ρ.

Repetition of this process for $\Delta_{d_0} < 0$ and for endpoints not on the $X_\rho Z_\rho$ $Z_\rho Y_\rho$ edges is straightforward. ■

D. Proof of Theorem 3

Proof: As stated in Section IV-C, the only parameters which achieve optimal rate of convergence place (d_1, d_2) along the $X_\rho Y_\rho$ edge of the ρ-convergence region, for any

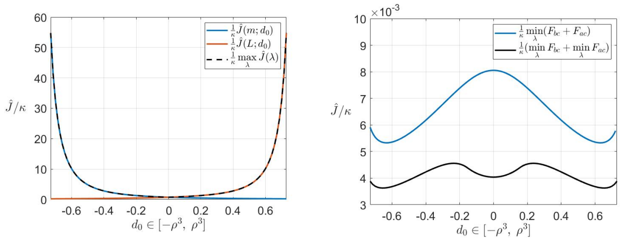


Fig. 4: Plots of $\hat{J}_{XY}(m; d_0)$ and $\hat{J}_{XY}(L; d_0)$ as functions of d_0, normalized by $\kappa = \frac{(1+\rho)^2}{(1-\rho)^2}$, for $\rho = 0.9$ for values of $d_0 \in [-\rho^3,\ \rho^3]$. The upper bound given in (30) is shown in black, showing the bound is achieved at every d_0. Figure (b) shows the true minimum of $\hat{J}(\lambda)$ over λ, computed numerically, in blue, while the lower bound given by the sum of (40) and (41) is shown in black.

fixed $d_0 \in [-\rho^3, \rho^3]$. We introduce the notation $\hat{J}_{XY}(\lambda; d_0)$ to refer to the modal contributions to variance amplification for these specific parameters.

The upper bound on $\hat{J}_{XY}(\lambda; d_0)$ has already been proven in Section IV-B. Notice that since our parameters place $(d_1, d_2)(L)$ at the Y_ρ vertex, the upper bound given in (30) is achieved.

The lower bound, similarly to the lower bound of Theorem 1, is obtained by bounding $F_{ac}(\lambda)$ and $F_{bc}(\lambda)$ independently, first for fixed d_0, and then for any d_0.

Proposition 2: Along the $X_\rho Y_\rho$ edge of $\Delta_\rho(d_0)$, with fixed d_0 F_{ac} achieves the minimum value

$$F_{ac} \ \geq \ \frac{(1-d_0)\rho^4}{2(1-\rho^2)(1+\rho)^2(d_0+\rho)^2(\hat{r}o^2-d_0)}. \qquad (40)$$

and F_{bc} is lower bounded by

$$F_{bc} \ \geq \ \frac{(2d_0\rho^4)}{(1+d_0)(1-\rho^2)(d_0+\rho^2)^3}. \qquad (41)$$

While minimizing the sum of (40) and (41) over d_0 is quite difficult, we can determine a loose lower bound by replacing d_0 with the maximum $d_0 = \rho^3$ and the minimum $d_0 = 0$ where appropriate, resulting in

$$\hat{J}_{XY}(\lambda) \ \geq \ \frac{1+\rho+\rho^2}{2(1+\rho)^5}. \qquad (42)$$

■

V. Concluding remarks

We study the class of three-step momentum algorithms that generalize heavy-ball and Nesterov's accelerated methods. For strongly convex quadratic problems, we have established algorithmic parameters which achieve the optimal convergence rate. Our results demonstrate that an additional momentum terms allowing increases the upper bound on modal contributions to variance amplification. Future work involves fully characterizing the Pareto-optimal curve of convergence rate and variance amplification for three-step momentum algorithms.

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] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imaging Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[6] L. Bottou and Y. Le Cun, "On-line learning for very large data sets," *Appl. Stoch. Models Bus. Ind.*, vol. 21, no. 2, pp. 137–151, 2005.

[7] M. Hong, M. Razaviyayn, Z.-Q. Luo, and J.-S. Pang, "A unified algorithmic framework for block-structured optimization involving big data: With applications in machine learning and signal processing," *IEEE Signal Process. Mag.*, vol. 33, no. 1, pp. 57–77, 2016.

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] Y. Bengio, "Gradient-based optimization of hyperparameters," *Neural computation*, vol. 12, no. 8, pp. 1889–1900, 2000.

[14] D. Maclaurin, D. Duvenaud, and R. Adams, "Gradient-based hyperparameter optimization through reversible learning," in *Proc. ICML*, 2015, pp. 2113–2122.

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

[16] 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.

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

[18] 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.

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

[20] 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.

[21] 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.

[22] H. Mohammadi, M. Razaviyayn, and M. R. Jovanović, "Tradeoffs between convergence rate and noise amplification for momentum-based accelerated optimization algorithms," 2022. [Online]. Available: https://arxiv.org/abs/2209.11920

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

[24] B. T. Polyak, "Comparison of the convergence rates for single-step and multi-step optimization algorithms in the presence of noise," *Engineering Cybernetics*, vol. 15, no. 1, pp. 6–10, 1977.

[25] E. I. Jury, "A simplified stability criterion for linear discrete systems," *Proceedings of the IRE*, vol. 50, no. 6, pp. 1493–1500, 1962.