IEEE Control Systems Letters paper presented at
2023 American Control Conference (ACC)
San Diego, CA, USA. May 31 - June 2, 2023

Noise amplifiation of momentum-based optimization algorithms

Hesameddin Mohammadi, Meisam Razaviyayn, and Mihailo R. Jovanović

Abstract— **We study momentum-based first-order optimization algorithms in which the iterations utilize information from the two previous steps and are subject to additive white noise. For strongly convex quadratic problems, we utilize Jury stability criterion to provide a novel geometric characterization of linear convergence and exploit this insight to derive alternative proofs of standard convergence results and identify fundamental performance tradeoffs. We use the steady-state variance of the error in the optimization variable to quantify noise amplification and establish analytical lower bounds on the product between the settling time and the smallest/largest achievable noise amplification that scale quadratically with the condition number. This extends the prior work [1], where only the special cases of Polyak's heavy-ball and Nesterov's accelerated algorithms were studied. We also use this geometric characterization to introduce a parameterized family of algorithms that strikes a balance between noise amplification and settling time while preserving order-wise Pareto optimality.**

Index Terms— **First-order methods, convergence rate, convex optimization, heavy-ball method, noise amplification, Nesterov's accelerated algorithm, performance tradeoffs, settling time.**

I. Introduction

Accelerated first-order algorithms [2]–[4] are often used for solving large-scale optimization problems [5]–[7] because of their fast convergence, low per-iteration complexity, and favorable scalability. Convergence properties of these algorithms have been carefully studied [8]–[14], but their performance in the presence of noise has received less attention [15]–[18]. Prior studies indicate that inaccuracies in gradient values can have larger negative impact on the convergence rate of accelerated methods compared to gradient descent [19]–[23].

Analyzing performance of accelerated algorithms with additive white noise that arises from uncertainty in gradient evaluation goes back to [24]; in this reference, Polyak established the optimal linear convergence rate for strongly convex quadratic problems and introduced time-varying parameters to achieve convergence in the error variance at a sub-linear rate but with an improved constant factor compared to gradient descent. With proper diminishing stepsize, acceleration in a sub-linear regime can also be achieved for general smooth convex problems [25]. For popular accelerated methods with constant parameters, control-theoretic tools were utilized in [1] to study the steady-state variance of the error in optimization variable for smooth strongly convex problems. For the parameters that optimize convergence rates for quadratic problems, tight upper and lower bounds on the noise amplification of gradient descent, heavy-ball method, and Nesterov's accelerated algorithms were developed [1]. These bounds are expressed in terms of the condition number κ and the problem dimension n, and they demonstrate opposite trends relative to the settling time: *for a fixed problem dimension n, accelerated algorithms increase noise amplification by a factor of $\Theta(\sqrt{\kappa})$ relative to gradient descent.* Furthermore, for strongly convex problems, tight and attainable upper bounds for noise amplification of gradient descent and Nesterov's accelerated method were provided [1].

In this paper, we extend the results of [1] to the class of first-order algorithms with three constant parameters in which the iterations involve information from the two previous steps. This class of algorithms includes heavy-ball and Nesterov's accelerated schemes as special cases and we examine its stochastic performance for strongly convex quadratic problems. Our results are complementary to [26], which evaluates stochastic performance in the objective error, and to recent work [27], which combines theoretical developments with computational experiments to demonstrate that a parameterized family of heavy-ball-like methods with reduced stepsize provides Pareto-optimal algorithms for simultaneous optimization of noise amplification and convergence rate. In contrast to [27], we establish an analytical lower bound on the product of the settling time and the noise amplification of two-step momentum method for any stabilizing algorithmic parameters. This lower bound scales with the square of the condition number and it reveals a fundamental limitation of this class of algorithms with constant stabilizing parameters.

Our results build upon a simple, yet powerful geometric viewpoint, which clarifies the relation between condition number, convergence rate, and algorithmic parameters for strongly convex quadratic problems. This allows us to present novel alternative proofs for (i) the optimal convergence rate of the two-step momentum algorithm, which recovers Nesterov's fundamental lower bound on the convergence rate [10]; and (ii) the optimal rates achieved by the special cases of standard gradient descent, heavy-ball method, and Nesterov's accelerated algorithms [11]. In addition, this viewpoint enables a novel geometric characterization of noise amplification in terms of stability margins and allows us to precisely quantify convergence/robustness tradeoffs.

II. Preliminaries and background

For unconstrained optimization problems

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

where $f\colon \mathbb{R}^n \to \mathbb{R}$ is a strongly convex function with a Lipschitz continuous gradient ∇f, we consider noisy

Financial support from the National Science Foundation under Awards ECCS 1708906 and ECCS 1809833 is gratefully acknowledged.

Hesameddin Mohammadi and Mihailo R. Jovanović are with the Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA 90089, USA. (e-mail: hesamedm@usc.edu; mihailo@usc.edu). Meisam Razaviyayn is with the Daniel J. Epstein Department of Industrial and Systems Engineering, University of Southern California, Los Angeles, CA 90089, USA (e-mail: razaviya@usc.edu).

979-8-3503-2806-6/$31.00 ©2023 AACC

momentum-based first-order algorithms in which iterations involve information from the two previous steps [18, Sec. 3],

$$x^{t+2} \;=\; x^{t+1} \,+\, \beta(x^{t+1} \,-\, x^t) \,- \\ \alpha \nabla f\big(x^{t+1} \,+\, \gamma(x^{t+1} \,-\, x^t)\big) \,+\, \sigma w^t. \tag{2-SM}$$

Here, t is the iteration index, x^t is the optimization variable, α is the stepsize, β and γ are momentum parameters used for acceleration, σ is the noise magnitude, and w^t is an additive white noise with zero mean and identity covariance matrix,

$$\mathbb{E}\left[w^t\right] \;=\; 0, \quad \mathbb{E}\left[w^t(w^\tau)^T\right] \;=\; I\,\delta(t-\tau)$$

where δ is the Kronecker delta and $\mathbb{E}$ is the expected value. If the only source of uncertainty is a noisy gradient, we set $\sigma = \alpha\sigma_a$ in (2-SM) to account for the effect of stepsize on the noise magnitude; otherwise we set $\sigma = \sigma_a$, where σ_a denotes the actual noise magnitude. Special cases of (2-SM) include noisy gradient descent ($\beta = \gamma = 0$), Polyak's heavy-ball ($\gamma = 0$), and Nesterov's accelerated algorithm ($\gamma = \beta$).

In the absence of noise (i.e., for $\sigma = 0$), the parameters (α, β, γ) can be selected such that the iterates converge linearly to the globally optimal solution $x^\star$ [10]. For the family of smooth strongly convex problems, the parameters that yield the fastest known linear convergence rate were provided in [13].

A. Linear dynamics for quadratic problems

Let $\mathcal{Q}_m^L$ denote the class of m-strongly convex L-smooth quadratic functions

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

with the condition number $\kappa := L/m$, where q is a vector and $Q = Q^T \succ 0$ is the Hessian matrix with eigenvalues

$$L \;=\; \lambda_1 \;\geq\; \lambda_2 \;\geq\; \ldots \;\geq\; \lambda_n \;=\; m \;>\; 0.$$

For $f \in \mathcal{Q}_m^L$, the linear time-invariant (LTI) state-space model

$$\psi^{t+1} \;=\; A\,\psi^t \;+\; B\,w^t, \quad z^t \;=\; C\,\psi^t \tag{3a}$$

with the state $\psi^t := [(x^t - x^\star)^T \; (x^{t+1} - x^\star)^T]^T$ and

$$A \;=\; \left[\begin{array}{cc} 0 & I \\ -\beta I + \gamma\alpha Q & (1+\beta)I - (1+\gamma)\alpha Q \end{array}\right] \tag{3b}$$
$$B^T \;=\; \left[\begin{array}{cc} 0 & \sigma I \end{array}\right], \quad C \;=\; \left[\begin{array}{cc} I & 0 \end{array}\right].$$

describes the *two-step momentum algorithm* (2-SM) with constant parameters (α, β, γ). Here, $z^t := x^t - x^\star$ is the performance output, and w^t is the white stochastic input.

B. Convergence rates

We call an algorithm stable if in the absence of noise (i.e., $\sigma = 0$), the state converges linearly with some rate $\rho < 1$,

$$\|\psi^t\| \;\leq\; c\,\rho^t\,\|\psi^0\| \;\text{ for all } t \;\geq\; 1 \tag{4}$$

for all functions $f \in \mathcal{Q}_m^L$, where $c > 0$ is a constant. For LTI system (3a), the spectral radius $\rho(A)$ of the matrix A determines the best achievable convergence rate.

For the class $\mathcal{Q}_m^L$ of infinite dimensional functions (i.e., for $n = \infty$), Nesterov established the fundamental lower bound on the convergence rate of any first-order algorithm [10],

$$1/(1 \,-\, \rho) \;\geq\; (\sqrt{\kappa} \,+\, 1)/2. \tag{5}$$

The quantity $1/(1-\rho)$ determines the *settling time*, i.e., the number of iterations required to reach a given desired accuracy [28, Appendix A]. This lower bound is sharp and it is achieved by the heavy-ball method (see Table I).

C. Noise amplification

For LTI system (3a) driven by an additive white noise w^t, $\mathbb{E}\left(\psi^{t+1}\right) = A\,\mathbb{E}\left(\psi^t\right)$. Thus, $\mathbb{E}\left(\psi^t\right) = A^t\,\mathbb{E}\left(\psi^0\right)$ and, for any stabilizing parameters (α, β, γ), the iterates reach a statistical steady-state with $\lim_{t\,\to\,\infty} \mathbb{E}\left(\psi^t\right) = 0$ and a variance that can be computed from the solution of the algebraic Lyapunov equation [1], [29]. We call the steady-state variance of the error noise (or variance) amplification,

$$J \;:=\; \lim_{t\,\to\,\infty} \frac{1}{t} \sum_{k\,=\,0}^{t} \mathbb{E}\left(\|x^k \,-\, x^\star\|^2\right). \tag{6}$$

Remark 1: An alternative performance metric that examines the steady-state variance of $y^t - x^\star$ was considered in [27], where $y^t := x^t + \gamma(x^t - x^{t-1})$ is the point at which the gradient is evaluated in (2-SM). Since for all $\gamma \geq 0$, we have $J_x \leq J_y \leq (1 + 2|\gamma|)^2 J_x$, where the subscripts x and y denote the noise amplification in terms of the error in x^t and y^t, these performance metrics are within a constant factor of each other for bounded values of γ.

D. Parameters that optimize convergence rate

For special instances of (2-SM), namely gradient descent (gd), heavy-ball method (hb), and Nesterov's accelerated algorithm (na), the parameters that optimize the convergence rates are given in [11, Proposition 1]. These parameters along with the corresponding rates and the noise amplification bounds are provided in Table I. The convergence rates are determined by the spectral radius of the corresponding A-matrices and the noise amplification bounds are computed by examining the solution to the algebraic Lyapunov equation and determining the functions $f \in \mathcal{Q}_m^L$ for which the steady-state variance is maximized/minimized [1].

For the parameters provided in Table I, there is a $\Theta(\sqrt{\kappa})$ improvement in settling times of the heavy-ball and Nesterov's accelerated algorithms relative to gradient descent,

$$1/(1 \,-\, \rho_{\rm gd}) \;=\; \Theta(\kappa), \quad 1/(1 \,-\, \rho_{{\rm hb},na}) \;=\; \Theta(\sqrt{\kappa}) \tag{7}$$

where $a = \Theta(b)$ means that a lies within constant factors of b as $b \to \infty$. In contrast to the convergence rate, the entire spectrum of Q influences noise amplification and its smallest and largest values

$$J_{\min} \;:=\; \min_{f\,\in\,\mathcal{Q}_m^L} J, \quad J_{\max} \;:=\; \max_{f\,\in\,\mathcal{Q}_m^L} J \tag{8}$$

over $\mathcal{Q}_m^L$ depend on the noise magnitude σ, the algorithmic parameters (α, β, γ), the problem dimension n, and the extreme eigenvalues m and L of Q. For the parameters that optimize convergence rates, tight upper and lower bounds on the noise amplification were developed in [1, Theorem 4]. These bounds are expressed in terms of the condition number κ and the problem dimension n, and they demonstrate

method	optimal parameters	$1/(1-\rho)$	$J_{\min}/\sigma^2$	$J_{\max}/\sigma^2$
Gradient	$\alpha = 2/(L+m)$, $\beta = \gamma = 0$	$(\kappa+1)/2$	$\Theta(\kappa) + n$	$n\Theta(\kappa)$
Heavy-ball	$\alpha = 4/(\sqrt{L} + \sqrt{m})^2$, $\beta = (1 - 2/(\sqrt{\kappa}+1))^2$, $\gamma = 0$	$(\sqrt{\kappa}+1)/2$	$\Theta(\kappa\sqrt{\kappa}) + n\Theta(\sqrt{\kappa})$	$n\Theta(\kappa\sqrt{\kappa})$
Nesterov	$\alpha = 4/(3L+m)$, $\beta = \gamma = 1 - 4/(\sqrt{3\kappa+1}+2)$	$\sqrt{3\kappa+1}/2$	$\Theta(\kappa\sqrt{\kappa}) + n$	$n\Theta(\kappa\sqrt{\kappa})$

TABLE I

SETTLING TIMES $1/(1-\rho)$ [11, PROPOSITION 1] ALONG WITH THE CORRESPONDING NOISE AMPLIFICATION (8) [1, THEOREM 4] FOR THE PARAMETERS THAT OPTIMIZE THE LINEAR CONVERGENCE RATE ρ FOR STRONGLY CONVEX QUADRATIC FUNCTION $f \in \mathcal{Q}_m^L$ WITH THE CONDITION NUMBER $\kappa := L/m$. HERE, n IS THE PROBLEM DIMENSION ($x \in \mathbb{R}^n$) AND σ^2 IS THE VARIANCE OF THE WHITE NOISE.

opposite trends relative to the settling time. In particular, for gradient descent,

$$J_{\min} \;=\; \sigma^2(\Theta(\kappa) + n), \;\; J_{\max} \;=\; \sigma^2 n\Theta(\kappa) \qquad \text{(9a)}$$

and for accelerated algorithms,

$$\begin{aligned} J_{\min} \;&=\; \begin{cases} \sigma^2(\Theta(\kappa\sqrt{\kappa}) \,+\, n\Theta(\sqrt{\kappa})) & \text{hb} \\ \quad\;\; \sigma^2(\Theta(\kappa\sqrt{\kappa}) \,+\, n) & \text{na} \end{cases} \qquad \text{(9b)} \\ J_{\max} \;&=\; \sigma^2 n\Theta(\kappa\sqrt{\kappa}). \end{aligned}$$

Thus, for fixed problem size n, accelerated algorithms increase noise amplification by a factor of $\Theta(\sqrt{\kappa})$ relative to gradient descent for the parameters that optimize convergence rates. While similar result also holds for heavy-ball and Nesterov's algorithms with arbitrary values of parameters α and β that provide convergence rate $\rho \leq 1 - c/\sqrt{\kappa}$ with $c > 0$ [1, Theorem 8], in this paper we establish the existence of a fundamental tradeoff between noise amplification and settling time for the two-step momentum method (2-SM) with arbitrary stabilizing values of constant parameters (α, β, γ).

III. MAIN RESULTS

In this section, we summarize our key contributions regarding robustness/convergence tradeoff for noisy (2-SM). A novel geometric characterization of conditions for stability and linear convergence allows us to provide alternative proofs of standard convergence results and quantify fundamental performance tradeoffs. We first provide an upper bound on noise amplification J in terms of the stability margin $1-\rho$ and then derive upper and lower bounds on the best achievable values of $J_{\min}/(1-\rho)$ and $J_{\max}/(1-\rho)$ given by (8).

A. Bounded noise amplification for stabilizing parameters

For a discrete-time LTI system with a convergence rate ρ, the distance of the eigenvalues to the unit circle is larger than $1-\rho$. We use this stability margin to establish an upper bound on the noise amplification J of the two-step momentum method (2-SM) for *any* stabilizing parameters (α, β, γ).

Theorem 1: Let (α, β, γ) be such that (2-SM) converges linearly with the rate $\rho < 1$ for all $f \in \mathcal{Q}_m^L$. Then,

$$J \;\leq\; \sigma^2 n(1 \,+\, \rho^2)/\left((1 \,+\, \rho)^3(1 \,-\, \rho)^3\right) \qquad \text{(10a)}$$

where n is the problem dimension. Furthermore, if $\sigma = \alpha\sigma_a$, i.e., when the only source of uncertainty is a noisy gradient,

$$J \;\leq\; \sigma_a^2 n(1 \,+\, \rho)(1 \,+\, \rho^2)/\left(L^2(1 \,-\, \rho)^3\right). \qquad \text{(10b)}$$

For $\rho < 1$, the bounds in (10) are increasing in ρ and become unbounded as $\rho \to 1$. In addition, both bounds are *exact* and they can be achieved by the heavy-ball method with the parameters that optimize the convergence rate (see Table I). We note that bounds in (10) are not tight for all stabilizing parameters. For example, applying (10a) to gradient descent with the optimal stepsize $\alpha = 2/(L+m)$ yields $J \leq \sigma^2 n\Theta(\kappa^3)$; this bound is off by a factor of κ^2; cf. Table I.

B. Tradeoff between settling time and noise amplification

For a fixed condition number κ and a problem size n, we are interested in designing parameters (α, β, γ) to simultaneously minimize settling time $1/(1-\rho)$ and noise amplification J. As noted in Section II, such a design may involve a tradeoff between these quantities. Since J depends on the entire spectrum of the Hessian matrix Q, we study this tradeoff by examining the smallest and largest values of J over the function class $\mathcal{Q}_m^L$ (i.e., $J_{\min}$ and $J_{\max}$ defined in (8)). We prove that the Pareto front for minimizing either $J_{\min}$ or $J_{\max}$ vs the settling time $1/(1-\rho)$ can be characterized by explicit upper and lower bounds on $J_{\min}/(1-\rho)$ and $J_{\max}/(1-\rho)$. These bounds scale quadratically with κ.

Theorem 2: Let (α, β, γ) be such that (2-SM) converges linearly with the rate $\rho < 1$ for all $f \in \mathcal{Q}_m^L$. Then,

$$\begin{aligned} J_{\min}/(1-\rho) \;&\geq\; \sigma^2\big(\kappa^2/64 \,+\, (n \,-\, 1)(\sqrt{\kappa}+1)/2\big) \\ J_{\max}/(1-\rho) \;&\geq\; \sigma^2\big((n \,-\, 1)\kappa^2/64 \,+\, (\sqrt{\kappa}+1)/2\big). \end{aligned}$$

where $J_{\min}, J_{\max}$ are given by (8). Furthermore, for $\sigma = \alpha\sigma_a$, i.e., when the only source of uncertainty is a noisy gradient,

$$\begin{aligned} J_{\min}/(1 \,-\, \rho) \;&\geq\; \sigma_a^2(\kappa^2 \,+\, (n \,-\, 1))/(2L)^2 \\ J_{\max}/(1 \,-\, \rho) \;&\geq\; \sigma_a^2((n \,-\, 1)\kappa^2 \,+\, 1)/(2L)^2. \end{aligned}$$

Since the above lower bounds hold for any stabilizing parameters, they also hold for the Pareto fronts that are obtained by minimizing $J_{\min}$ and $J_{\max}$ over (α, β, γ). These bounds scale as $\Theta(\kappa^2)$ when the problem dimension n is fixed. Theorem 3 establishes $\Theta(\kappa^2)$ upper bounds and demonstrates the tightness of this scaling.

Theorem 3: For the class of functions $\mathcal{Q}_m^L$ with condition number $\kappa = L/m$, let the scalar ρ be such that

$$1/(1-\rho) \;\in\; [(\sqrt{\kappa}+1)/2, (\kappa+1)/2].$$

Then, the two-step momentum algorithm (2-SM) with

$$\alpha \;=\; (1 \,+\, \rho)(1 \,+\, c\rho)/L, \;\; \beta \;=\; c\rho^2, \;\; \gamma \;=\; 0 \qquad \text{(11)}$$

achieves the settling time $1/(1-\rho)$ and satisfies

$$J_{\min}/(1\,-\,\rho)\;\leq\;\sigma^2\kappa\,(2\kappa\,+\,n) \tag{12a}$$

$$J_{\max}/(1\,-\,\rho)\;\leq\;\sigma^2 n\kappa(\kappa\,+\,1) \tag{12b}$$

where $J_{\min},\ J_{\max}$ are given by (8) and $c\in[0,\,1]$ satisfies

$$c\;=\;\left((1-\rho)\kappa-(1+\rho)\right)/\left(\rho\left((1-\rho)\kappa+(1+\rho)\right)\right).$$

Theorem 3 establishes $\Theta(\kappa^2)$ upper bounds on $J_{\min}/(1-\rho)$ and $J_{\max}/(1-\rho)$ for a parameterized family of heavy-ball-like algorithms with $\gamma=0$. In these upper bounds, the problem dimension n appears in an additive fashion in (12a) and in a multiplicative fashion in (12b). In addition, the boundaries of the interval for $1/(1-\rho)$ are determined by the settling times of the heavy-ball method and gradient descent with parameters given in Table I. The lower and upper bounds in Theorems 2 and 3 scale as $\Theta(\kappa^2)$ and they are order-wise tight.

IV. Geometric characterization

In this section, we examine the relation between the convergence rate and noise amplification of (2-SM) for $\mathcal{Q}_m^L$. In particular, we use the eigenvalue decomposition of the Hessian matrix Q to bring the dynamics into n decoupled second-order systems. We utilize Jury stability criterion to provide novel geometric characterization of stability and ρ-linear convergence and derive alternative proofs of standard convergence results and quantify fundamental performance tradeoffs.

A. Modal decomposition

We utilize the eigenvalue decomposition of the Hessian matrix $Q=V\Lambda V^T$, where Λ is the diagonal matrix of the eigenvalues and V is the orthogonal matrix of the corresponding eigenvectors. The change of variables $\hat{x}^t:=V^T(x^t-x^\star)$ and $\hat{w}^t:=V^Tw^t$ allows us to bring system (3) into n subsystems,

$$\hat{\psi}_i^{t+1}\;=\;\hat{A}_i\hat{\psi}_i^t\,+\,\hat{B}_i\hat{w}_i^t,\quad \hat{z}_i^t\;=\;\hat{C}_i\hat{\psi}_i^t \tag{13a}$$

where $\hat{\psi}_i^t=\left[\begin{array}{cc}\hat{x}_i^t & \hat{x}_i^{t+1}\end{array}\right]^T$ is the state and

$$\begin{array}{rcl}\hat{A}_i &=& \hat{A}(\lambda_i)\;:=\;\left[\begin{array}{cc}0 & 1\\ -a(\lambda_i) & -b(\lambda_i)\end{array}\right]\\ \hat{B}_i &=& \left[\begin{array}{cc}0 & \sigma\end{array}\right]^T,\quad \hat{C}_i\;=\;\left[\begin{array}{cc}1 & 0\end{array}\right]\end{array} \tag{13b}$$

$$a(\lambda):=\beta-\gamma\alpha\lambda,\ b(\lambda):=(1+\gamma)\alpha\lambda-(1+\beta). \tag{13c}$$

B. Conditions for linear convergence

The convergence rate ρ for $\mathcal{Q}_m^L$ is determined by

$$\rho\;=\;\max_{\lambda\,\in\,[m,L]}\;\rho(\hat{A}(\lambda)). \tag{14}$$

For the heavy-ball and Nesterov's accelerated methods, analytical expressions for the spectral radius $\rho(\hat{A}(\lambda))$ were developed and algorithmic parameters that optimize convergences rate were obtained in [11]. Unfortunately, these expressions do not provide insight into the relation between convergence rates and noise amplification.

In this paper, we ask the dual question:

- *For a fixed convergence rate ρ, what is the largest condition number κ that can be handled by* (2-SM) *with constant parameters* (α,β,γ)?

We note that the matrices $\hat{A}(\lambda)$ share the same structure as

$$M\;=\;\left[\begin{array}{cc}0 & 1\\ -a & -b\end{array}\right] \tag{15a}$$

with the real scalars a and b and that the characteristic polynomial associated with the matrix M is given by

$$F(z)\;:=\;\det\left(zI\,-\,M\right)\;=\;z^2\,+\,b\,z\,+\,a. \tag{15b}$$

We next utilize the Jury stability criterion [30, Chap. 4-3] to provide conditions for stability of the matrix M given by (15a).

Lemma 1: For the matrix $M\in\mathbb{R}^{2\times 2}$ in (15a), we have $\rho(M)<1$ if and only if $(b,a)\in\Delta$, where the stability set

$$\Delta\;:=\;\{(b,a)\mid|b|\,-\,1\,<\,a\,<\,1\} \tag{16a}$$

is an open triangle in the (b,a)-plane with vertices

$$X\;=\;(-2,1),\quad Y\;=\;(2,1),\quad Z\;=\;(0,-1). \tag{16b}$$

For any $\rho>0$, the spectral radius $\rho(M)$ of the matrix M is smaller than ρ if and only if $\rho(M/\rho)$ is smaller than 1. This observation in conjunction with Lemma 1 allow us to obtain necessary and sufficient conditions for stability with the linear convergence rate ρ of (2-SM).

Lemma 2: For any positive scalar $\rho<1$ and the matrix $M\in\mathbb{R}^{2\times 2}$ in (15a), we have $\rho(M)\leq\rho$ if and only if $(b,a)\in\Delta_\rho$, where the ρ-linear convergence set

$$\Delta_\rho\;:=\;\left\{(b,a)\mid\rho\left(|b|\,-\,\rho\right)\,\leq\,a\,\leq\,\rho^2\right\} \tag{17a}$$

is a closed triangle in the (b,a)-plane with vertices

$$X_\rho\;=\;(-2\rho,\rho^2),\ Y_\rho\;=\;(2\rho,\rho^2),\ Z_\rho\;=\;(0,-\rho^2). \tag{17b}$$

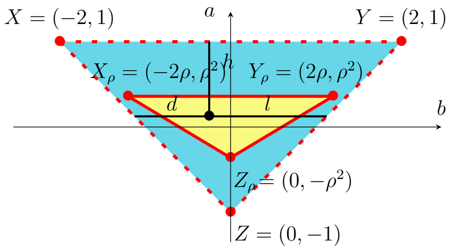

Fig. 1. The stability set Δ (the open, cyan triangle) in (16a) and the ρ-linear convergence set Δ_ρ (the closed, yellow triangle) in (17a). For the point (b,a) (black bullet) associated with the matrix M in (15a), the corresponding distances (d,h,l) in (22) are marked by black lines.

Figure 1 shows the stability and the ρ-linear convergence sets Δ and Δ_ρ. We use Lemmas 1, 2 along with the fact that $a(\lambda)$ and $b(\lambda)$ in (13c) satisfy the affine relation

$$(1\,+\,\gamma)a(\lambda)\,+\,\gamma b(\lambda)\;=\;\beta-\gamma \tag{18}$$

to characterize the convergence rate of (2-SM).

Lemma 3: The two-step momentum algorithm (2-SM) with constant parameters (α,β,γ) is stable for all functions $f\in$

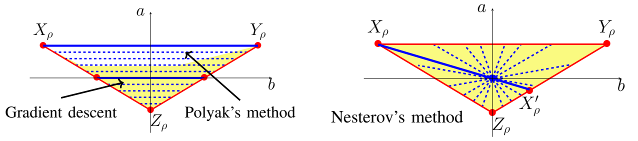

Fig. 2. For a fixed ρ-linear convergence triangle Δ_ρ (yellow), dashed blue lines mark the line segments $(b(\lambda), a(\lambda))$ with $\lambda \in [m, L]$ for gradient descent, Polyak's heavy-ball, and Nesterov's accelerated methods. The solid blue line segments correspond to the parameters for which the algorithm achieves rate ρ for the largest possible condition number.

$\mathcal{Q}_m^L$ if and only if the following equivalent conditions hold:

1) $(b(\lambda), a(\lambda)) \in \Delta$ for all $\lambda \in [m, L]$;
2) $(b(\lambda), a(\lambda)) \in \Delta$ for $\lambda = \{m, L\}$.

Furthermore, the linear convergence rate $\rho < 1$ is achieved for all functions $f \in \mathcal{Q}_m^L$ if and only if the following equivalent conditions hold:

1) $(b(\lambda), a(\lambda)) \in \Delta_\rho$ for all $\lambda \in [m, L]$,
2) $(b(\lambda), a(\lambda)) \in \Delta_\rho$ for $\lambda = \{m, L\}$.

Here, $(b(\lambda), a(\lambda))$ is given by (13c), and the sets Δ and Δ_ρ are given by (16a) and (17a), respectively.

Lemma 3 exploits the affine relation (18) between $a(\lambda)$ and $b(\lambda)$ and the convexity of the sets Δ and Δ_ρ to establish necessary and sufficient conditions for stability and ρ-linear convergence: the inclusion of the end points of the line segment $(b(\lambda), a(\lambda))$ associated with the extreme eigenvalues m and L of the matrix Q in the corresponding triangle. A similar approach was taken in [27, Appendix A.1], where the affine nature of the conditions resulting from the Jury stability criterion with respect to λ was used to conclude that $\rho(\hat{A}(\lambda))$ is a quasi-convex function of λ and show that the extreme points m and L determine $\rho(A)$. In contrast, we exploit the triangular shapes of the stability and ρ-linear convergence sets and utilize this geometric insight to identify the parameters that optimize the convergence rate and to establish tradeoffs between noise amplification and convergence rate.

For the two-step momentum algorithm (2-SM) with constant parameters, Lemma 3 provides a simple alternative proof for the fundamental lower bound (5) on the settling time established by Nesterov. Our proof utilizes the fact that for any point $(b(\lambda), a(\lambda)) \in \Delta_\rho$, the horizontal signed distance to the edge XZ of the stability triangle Δ (see Figure 1) satisfies

$$d(\lambda) \; := \; a(\lambda) \, + \, b(\lambda) \, + \, 1 \; = \; \alpha\lambda. \qquad (19)$$

Proposition 1: Let (2-SM) achieve the linear convergence rate $\rho < 1$ for all functions $f \in \mathcal{Q}_m^L$. Then, lower bound (5) on the settling time holds and it is achieved by the heavy-ball method with the parameters provided in Table I.

Proof: Let $d(m) = \alpha m$ and $d(L) = \alpha L$ denote the values of the function $d(\lambda)$ associated with the points $(b(m), a(m))$ and $(b(L), a(L))$, where (b, a) and d are given by (13c) and (19), respectively. Lemma 3 implies that $(b(L), a(L))$ and $(b(m), a(m))$ lie in the ρ-linear convergence triangle Δ_ρ. Thus, $d_{\max}/d_{\min} \geq d(L)/d(m) = \kappa$, where $d_{\max}$ and $d_{\min}$ denote the largest and smallest values that d can take among all points $(b, a) \in \Delta_\rho$. From the shape of Δ_ρ, we conclude that $d_{\max}$ and $d_{\min}$ correspond to the vertices Y_ρ and X_ρ of Δ_ρ given by (17b); see Figure 1. Thus,

$$d_{\max} \; = \; d_{Y_\rho} \; = \; 1 \, + \, \rho^2 \, + \, 2\,\rho \; = \; (1 \, + \, \rho)^2 \qquad (20a)$$
$$d_{\min} \; = \; d_{X_\rho} \; = \; 1 \, + \, \rho^2 \, - \, 2\,\rho \; = \; (1 \, - \, \rho)^2. \qquad (20b)$$

This in conjunction with the previous inequality yields

$$\kappa \; = \; d(L)/d(m) \; \leq \; d_{\max}/d_{\min} \; = \; (1+\rho)^2/(1-\rho)^2. \qquad (21)$$

Rearranging terms above gives lower bound (5). ■

To provide insight, we next examine the implications of Lemma 3 for gradient descent, Polyak's heavy-ball, and Nesterov's accelerated algorithms. In all three cases, our dual approach recovers the optimal convergence rates provided in Table I. From the definition of a and b in (13c), it follows that the line segment $(b(\lambda), a(\lambda))$ with $\lambda \in [m, L]$ satisfies:

- gradient descent ($\beta = \gamma = 0$): (b, a) is the horizontal line segment parameterized by $a = 0$;
- heavy-ball method ($\gamma = 0$): (b, a) is the horizontal line segment parameterized by $a = \beta$; and
- Nesterov's accelerated method ($\beta = \gamma$): (b, a) is the line segment parameterized by $a = -\beta b/(1 + \beta)$.

These observations are illustrated in Figure 2. To obtain the largest possible condition number for which the convergence rate ρ is feasible for each algorithm, one needs to find the largest ratio $d(L)/d(m) = \kappa$ among possible orientations for the line segment $(b(\lambda), a(\lambda))$ with $\lambda \in [m, L]$ to lie in Δ_ρ.

- For gradient descent, the largest ratio $d(L)/d(m)$ corresponds to the intersections of the horizontal axis and the edges $Y_\rho Z_\rho$ and $X_\rho Z_\rho$ of the triangle Δ_ρ, which are given by $(\rho, 0)$ and $(-\rho, 0)$, respectively. Thus,

  $$\kappa \; = \; d(L)/d(m) \; \leq \; (1 \, + \, \rho)/(1 \, - \, \rho).$$

  Rearranging terms above yields a lower bound on the settling time $1/(1-\rho) \geq (\kappa+1)/2$. This lower bound is tight as it can be achieved by choosing the parameters in Table I, which place $(b(\lambda), a(\lambda))$ to $(\rho, 0)$ and $(-\rho, 0)$ for $\lambda = L$ and $\lambda = m$, respectively.
- For the heavy-ball method, the optimal rate is recovered by designing the parameters (α, β) such that the vertices X_ρ and Y_ρ belong to the line segment $(b(\lambda), a(\lambda))$,

  $$\kappa \; = \; d(L)/d(m) \; \leq \; (1 \, + \, \rho)^2/(1 \, - \, \rho)^2.$$

  By choosing $d(L) = d_{Y_\rho}$ and $d(m) = d_{X_\rho}$, we recover the optimal parameters provided in Table I and achieve the lower bound (5) on the convergence rate.
- For Nesterov's accelerated method, we can verify that the largest ratio $d(L)/d(m)$ corresponds to the line segment $X_\rho X'_\rho$ that passes through the origin, where $X'_\rho = (2\rho/3, -\rho^2/3)$ lies on the edge $Y_\rho Z_\rho$ This yields

  $$\kappa \; = \; d(L)/d(m) \; \leq \; (1 \, + \, 2\rho/3 \, - \, \rho^2/3)/(1 \, - \, \rho)^2.$$

  Rearranging terms in this inequality provides a lower bound on the settling time $1/(1-\rho) \geq \sqrt{3\kappa + 1}/2$. This lower bound is tight and it can be achieved with the parameters provided in Table I, which place $(b(L), a(L))$ to X'_ρ and $(b(m), a(m))$ to X_ρ; see Figure 2.

C. Noise amplification

To quantify the noise amplification of the two-step momentum algorithm (2-SM), we utilize an alternative characterization of the stability and ρ-linear convergence triangles Δ and Δ_ρ. Let d and l denote the horizontal signed distances of the point (a, b) to the edges XZ and YZ of Δ (see Figure 1),

$$\begin{aligned} d(\lambda) &:= a(\lambda) + b(\lambda) + 1 \\ l(\lambda) &:= a(\lambda) - b(\lambda) + 1. \end{aligned} \tag{22a}$$

and let h denote its vertical signed distance to the edge XY,

$$h(\lambda) := 1 - a(\lambda). \tag{22b}$$

Then, from the definition of the set Δ, we have $(b, a) \in \Delta$ if and only if $h,\, d,\, l > 0$. In Theorem 4, we quantify the steady-state variance of the error in the optimization variable in terms of the spectrum of the Hessian matrix and the algorithmic parameters for noisy two-step momentum algorithm (2-SM).

Theorem 4: For $f \in \mathcal{Q}_m^L$ with the Hessian matrix Q, the steady-state variance of $x^t - x^\star$ for (2-SM) with any stabilizing parameters (α, β, γ) is determined by

$$J \;=\; \sum_{i=1}^{n} \frac{\sigma^2(d(\lambda_i) + l(\lambda_i))}{2\, d(\lambda_i)\, h(\lambda_i)\, l(\lambda_i)} \;=:\; \sum_{i=1}^{n} \hat{J}(\lambda_i)$$

Here, $\hat{J}(\lambda_i)$ denotes the modal contribution of the ith eigenvalue λ_i of Q to the steady-state variance, (d, h, l) are defined in (22), and (a, b) are given by (13c).

V. Concluding remarks

We studied two-step momentum algorithms subject to additive white noise and established lower bounds on the product of the settling time and the smallest/largest noise amplification. These bounds scale as κ^2 for all stabilizing parameters and they reveal a fundamental limitation imposed by the condition number in designing algorithms that tradeoff noise amplification and convergence rate. Our analysis uses a novel geometric viewpoint on the relation between noise amplification, convergence rate, and algorithmic parameters, and provides an alternative proof for optimal convergence rates of the heavy-ball and Nesterov's accelerated methods. Our ongoing work includes extending these results to algorithms with more complex structures that involve information from more than the last two iterates and time varying parameters.

References

[1] 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.

[2] 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.

[3] Y. Nesterov, "A method for solving the convex programming problem with convergence rate $O(1/k^2)$," in *Dokl. Akad. Nauk SSSR*, vol. 27, 1983, pp. 543–547.

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

[5] 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.

[6] 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.

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

[11] 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.

[12] S. Cyrus, B. Hu, B. Van Scoy, and L. Lessard, "A robust accelerated optimization algorithm for strongly convex functions," in *Proceedings of the 2018 American Control Conference*, 2018, pp. 1376–1381.

[13] 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.

[14] M. Fazlyab, A. Ribeiro, M. Morari, and V. M. Preciado, "Analysis of optimization algorithms via integral quadratic constraints: Nonstrongly convex problems," *SIAM J. Optim.*, vol. 28, no. 3, pp. 2654–2689, 2018.

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

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

[17] A. Beirami, M. Razaviyayn, S. Shahrampour, and V. Tarokh, "On optimal generalizability in parametric learning," in *Proc. Neural Information Processing (NIPS)*, 2017, pp. 3458–3468.

[18] K. Yuan, B. Ying, and A. H. Sayed, "On the influence of momentum acceleration on online learning," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 6602–6667, 2016.

[19] 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.

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

[21] 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.

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

[23] 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.

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

[25] O. Devolder, "Stochastic first order methods in smooth convex optimization," Catholic Univ. Louvain, Louvain-la-Neuve, Tech. Rep., 2011.

[26] N. S. Aybat, A. Fallah, M. M. Gürbüzbalaban, and A. Ozdaglar, "Robust accelerated gradient methods for smooth strongly convex functions," *SIAM J. Opt.*, vol. 30, no. 1, pp. 717–751, 2020.

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

[28] H. Mohammadi, M. Razaviyayn, and M. R. Jovanović, "Tradeoffs between convergence rate and noise amplification for momentum-based accelerated optimization algorithms," 2022, arXiv:2209.11920v1.

[29] H. Kwakernaak and R. Sivan, *Linear optimal control systems*. Wiley-Interscience, 1972.

[30] K. Ogata, *Discrete-time control systems*. New Jersey: Prentice-Hall, 1994.