2019 American Control Conference (ACC)
Philadelphia, PA, USA, July 10-12, 2019

Performance of noisy Nesterov's accelerated method for strongly convex optimization problems

Hesameddin Mohammadi, Meisam Razaviyayn, and Mihailo R. Jovanović

***Abstract*— We study the performance of noisy gradient descent and Nesterov's accelerated methods for strongly convex objective functions with Lipschitz continuous gradients. The steady-state second-order moment of the error in the iterates is analyzed when the gradient is perturbed by an additive white noise with zero mean and identity covariance. For any given condition number κ, we derive explicit upper bounds on noise amplification that only depend on κ and the problem size. We use quadratic objective functions to derive lower bounds and to demonstrate that the upper bounds are tight up to a constant factor. The established upper bound for Nesterov's accelerated method is larger than the upper bound for gradient descent by a factor of $\sqrt{\kappa}$. This gap identifies a fundamental tradeoff that comes with acceleration in the presence of stochastic uncertainties in the gradient evaluation.**

***Index Terms*— Accelerated first-order algorithms, control for optimization, convex optimization, integral quadratic constraints, linear matrix inequalities, Nesterov's method, noise amplification, second-order moments, semidefinite programming.**

I. Introduction

First-order methods are frequently used in solving modern large-scale optimization problems [1]–[3]. Gradient descent as well as its accelerated counterparts are popular due to their simplicity and scalability. These algorithms have been extensively studied under different stepsize selection rules [4]–[9].

In many applications, the exact values of the objective function and/or its gradient are not fully available. This happens when the objective function is obtained via costly simulations (e.g., tuning of hyper-parameters in supervised/unsupervised learning [10]–[12]), when the objective function is evaluated through noisy measurements (e.g., real-time and embedded applications), and when the computations are done over network [13]. Another application arises in (batch) stochastic gradient settings where at each iteration the gradient of the objective function is computed from a small batch of data points. Such a batch gradient is known to be a noisy unbiased estimator for the gradient of the training loss. Moreover, deliberately adding noise to the algorithm iterates may help escaping saddle points [14], [15].

In all of the above scenarios, the iterative algorithms only have access to noisy estimates of the gradient. This motivates the study of gradient descent and its accelerated variant in the presence of noisy/inexact gradient oracles [16]–[19]. These studies propose different strategies for dealing with the additive gradient noise. For example, while early stochastic approximation results suggest to use a stepsize that is inversely proportional to the iteration number [17], a more robust behavior can be obtained by combining larger stepsizes with averaging [18], [20]. Utility of these averaging schemes and their modifications for solving quadratic optimization and manifold problems has been examined in [21]–[23]. In addition, the convergence of (accelerated) gradient descent has been studied under the presence of noise. For example, an upper bound on the error in iterates for accelerated proximal gradient methods was established in [24] when the gradient is perturbed by a deterministic noise. Using this upper bound, it has been shown that both proximal gradient and accelerated proximal gradient can maintain their convergence rates provided that the error vanishes fast enough [24]. It has been also observed that accelerated first-order algorithms are more susceptible to noise than their non-accelerated variants [19], [24]–[27].

In this paper, we utilize techniques from control theory to study the effect of additive white noise on the performance of gradient descent and Nesterov's accelerated method. We confine our analysis to conventional algorithmic parameters for strongly convex objective functions with Lipschitz gradients. For quadratic problems, the linearity of the gradient allows for the steady-state second-order moment of the error to be explicitly computed as a function of eigenvalues of the Hessian of the objective function [28]. This characterization reveals that for quadratic problems with a condition number smaller than κ, acceleration may increase the steady-state second-order moment by a factor of $\sqrt{\kappa}$.

We extend this result from quadratic to general strongly convex problems. While exact characterization of the steady-state second-order moments for general problems is challenging because of the nonlinear dynamics, we show how upper bounds on this quantity can be obtained using a framework that utilizes concepts from robust control theory. This framework was first developed for the analysis of optimization algorithms in [29] and it has also been employed and further improved [30]–[33] to study convergence and robustness of the first-order methods to deterministic perturbations in the gradient. However, to the best of our knowledge, the analysis of second-order moments for algorithms with additive white stochastic disturbances has not been done before.

Amongst the class of smooth and strongly convex problems with condition number κ, we demonstrate that there exists

Financial support from the National Science Foundation under award ECCS-1809833 and from the Air Force Office of Scientific Research under award FA9550-16-1-0009 is gratefully acknowledged.

H. Mohammadi and M. R. Jovanović are with the Ming Hsieh Department of Electrical and Computer Engineering; M. Razaviyayn is with the Epstein Department of Industrial and Systems Engineering, University of Southern California, Los Angeles, CA 90089. E-mails: hesamedm@usc.edu, razaviya@usc.edu, mihailo@usc.edu.

978-1-5386-7928-9/$31.00 ©2019 AACC

a quadratic function for which the steady-state second-order moment of the error is most amplified (up to constant factors). For functions with the condition number smaller than κ, this shows that acceleration may increase noise amplification by up to a factor of $\sqrt{\kappa}$.

The rest of the paper is structured as follows. In Section II, we formulate the problem and provide background material used in the paper. We also outline our method for quantifying the steady-state second-order moment of the optimization variable. In Section III, we restrict our attention to the class of strongly convex quadratic problems and obtain lower bounds on the noise amplification. The LMI-based framework that provides upper bounds is presented in Section IV and the paper is concluded in Section V.

Notation: We write $f = \Omega(g)$ (or, equivalently, $g = O(f)$) to denote the existence of positive constants c_i such that, for any $x > c_2$, the functions f and g: $\mathbb{R} \to \mathbb{R}$ satisfy $f(x) \geq c_1 g(x)$. We write $f = \Theta(g)$ if both $f = \Omega(g)$ and $f = O(g)$. The smallest and largest eigenvalues of a matrix are denoted by $\lambda_{\min}$ and $\lambda_{\max}$.

II. Problem formulation and background

Consider the unconstrained optimization problem

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

where f: $\mathbb{R}^n \to \mathbb{R}$ is a smooth strongly convex function. We study two commonly used first-order methods for solving problem (1), gradient descent

$$x^{t+1} \;=\; x^t \;-\; \alpha \nabla f(x^t)$$

and Nesterov's accelerated method

$$\begin{aligned} x^{t+2} \;=\; x^{t+1} \;+\; \beta(x^{t+1} - x^t) \;- \\ \alpha \nabla f\big(x^{t+1} \;+\; \beta(x^{t+1} \;-\; x^t)\big) \,. \end{aligned}$$

Here, x^t is the optimization variable, α is the stepsize, and β is the extrapolation parameter. Clearly, gradient decent can be obtained by setting $\beta = 0$ in Nesterov's formulation.

We denote by $\mathcal{F}_m^L$ the set of functions f that are m-strongly convex and L-smooth; $f \in \mathcal{F}_m^L$ means that $f(x) - \frac{m}{2}\|x\|^2$ is convex and that the gradient ∇f is L-Lipschitz continuous. We associate with $\mathcal{F}_m^L$ the condition number $\kappa := L/m$. In particular, for a twice continuously differentiable function f with the Hessian $\nabla^2 f$, we have

$$f \,\in\, \mathcal{F}_m^L \;\Leftrightarrow\; mI \;\preceq\; \nabla^2 f(x) \;\preceq\; LI, \quad \forall x \,\in\, \mathbb{R}^n.$$

For $f \in \mathcal{F}_m^L$, the parameters α and β can be selected such that both gradient descent and Nesterov's accelerated method converge to the global minimum $x^\star$ of (1) at a linear rate ρ. Table I provides the conventional values of these parameters and the corresponding guaranteed convergence rates [9].

Method	Parameters	Rate bound
Gradient	$\alpha \,=\, \frac{1}{L}, \beta \,=\, 0$	$\rho \,\leq\, 1 \,-\, \frac{1}{2\kappa}$
Nesterov	$\alpha \,=\, \frac{1}{L}, \beta \,=\, \frac{\sqrt{\kappa}\,-\,1}{\sqrt{\kappa}\,+\,1}$	$\rho \,\leq\, 1 \,-\, \frac{1}{2\sqrt{\kappa}}$

TABLE I: Conventional values of parameters and the corresponding rate bounds for $f \in \mathcal{F}_m^L$ where $\kappa := L/m$.

We study the effect of stochastic uncertainties in gradient evaluation on the performance of gradient descent and Nesterov's accelerated method. In particular, we add a white stochastic process w^t with zero mean and identity covariance matrix (i.e., $\mathbb{E}(w^t) = 0$ and $\mathbb{E}(w^t(w^\tau)^T) = I\,\delta(t-\tau)$, where δ is the Kronecker delta and $\mathbb{E}$ is the expectation operator) to the iterates of Nesterov's accelerated algorithm,

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

and note that gradient descent can be identified as a special case of (2) with $\beta = 0$. For this noisy algorithm, we examine the steady-state second-order moment of the error $x^t - x^\star$,

$$J \;:=\; \lim_{t \,\to\, \infty} \mathbb{E}\left(\|x^t - x^\star\|^2\right)$$

as a metric to assess sensitivity of (2) to noise. We focus on algorithm (2) with parameters provided in Table I and demonstrate that although gradient descent does not have an optimal convergence rate, it outperforms Nesterov's method when it comes to the second-order moment analysis. In what follows, we use subscripts gd and na (e.g., J_{gd} and J_{na}) to denote quantities that correspond to gradient descent and Nesterov's method with the conventional values of parameters provided in Table I. Furthermore, without loss of generality we assume that $x^\star = 0$ is the unique minimizer of (1).

The second-order moment analysis of dynamical systems has been well studied in the controls literature. For stable linear time-invariant systems, the steady-state variance of the output can be directly computed from the solution of the algebraic Lyapunov equation. For nonlinear systems, although there is no explicit characterization for the noise amplification in general, methods from control theory can be utilized to find upper bounds.

Among all functions $f \in \mathcal{F}_m^L$, we define $J^\star$ to be the steady-state second-order moment of the optimization variable x in algorithm (2) with respect to the worst objective function,

$$J^\star \;:=\; \sup_{f \,\in\, \mathcal{F}_m^L} \; J. \tag{3}$$

The goal of this paper is to evaluate this quantity for both gradient descent $J_{\text{gd}}^\star$ and Nesterov's accelerated method $J_{\text{na}}^\star$ with parameters provided in Table I. However, since algorithm (2) is a nonlinear dynamical system, direct computation of $J^\star$ for general $f \in \mathcal{F}_m^L$ is challenging. In order to overcome this challenge, we take a two-step approach:

1) We exploit a result from [28] to establish a lower bound for $J^\star$ by restricting f to be a quadratic function.
2) We employ the theory of Integral Quadratic Constraints (IQCs) to determine upper bounds on $J^\star$ by solving certain Linear Matrix Inequalities (LMIs).

In the first step, we define q to be the largest steady-state second-order moment of the optimization variable x in algorithm (2) when the objective function is quadratic, i.e.,

$$q \,:=\, \sup_{f \,\in\, \mathcal{F}_m^L} \; J \qquad \text{subject to} \;\; f \text{ is quadratic.} \tag{4}$$

Clearly, $J^\star \geq q$ and therefore q can be used as a lower bound for $J^\star$. In the second step, small sizes of the resulting LMIs (2×2 for gradient descent and 3×3 for Nesterov's accelerated method) allow us to obtain explicit upper bounds. Comparing these upper bounds with the explicit lower bound q obtained using quadratic functions, we demonstrate that both upper and lower bounds are tight up to constant factors, thereby providing accurate approximations for $J^\star$.

III. Quadratic optimization problems

Let the objective function in optimization problem (1) be a strongly convex quadratic function,

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

where Q is a symmetric positive definite matrix. In this case, the linearity of the gradient mapping $\nabla f(x) = Qx$ allows us to cast algorithm (2) as a linear dynamical system

$$\begin{aligned} \psi^{t+1} \,&=\, A\,\psi^t \,+\, B\,w^t \\ z^t \,&=\, C\,\psi^t \end{aligned} \tag{5}$$

with the state $\psi^t := [\,(x^t)^T \;\; (x^{t+1})^T\,]^T$, the performance output z^t, the input w^t, and

$$A \,=\, \begin{bmatrix} 0 & I \\ -\beta(I - \alpha Q) & (1+\beta)(I - \alpha Q) \end{bmatrix}$$

$$B \,=\, \begin{bmatrix} 0 \\ I \end{bmatrix}, \quad C \,=\, \begin{bmatrix} I & 0 \end{bmatrix}.$$

For a white stochastic input w^t with zero mean and identity covariance matrix, the steady-state second-order moment J of the output z^t in (5) can be computed from the solution of the algebraic Lyapunov equation,

$$P \,=\, A P A^T \,+\, B B^T$$

as

$$J \,=\, \mathrm{trace}\,(C\,P\,C^T)$$

where $P := \lim_{t \to \infty} \mathbb{E}\,(\psi^t(\psi^t)^T)$ is the steady-state covariance matrix of the state vector ψ^t. Recently, this approach was utilized in [28] to obtain the following result.

Theorem 1: Let $f(x) := \frac{1}{2}\, x^T Q x$ with $Q = Q^T \succ 0$. The steady-state second-order moment J of the optimization variable $x^t \in \mathbb{R}^n$ for algorithm (2) with any constant stabilizing parameters α and β is given by $J = \sum_{i\,=\,1}^n \hat{J}(\lambda_i)$, where λ_i are the eigenvalues of Q and

$$\hat{J}(\lambda_i) := \frac{1 + \beta(1 - \alpha\lambda_i)}{\alpha\lambda_i\,(1 - \beta(1 - \alpha\lambda_i))\,(2(1+\beta) - (2\beta+1)\alpha\lambda_i)}.$$

Theorem 1 provides an explicit characterization of the steady-state second-order moment J for strongly convex quadratic objective functions. This characterization reveals that J is influenced by the entire spectrum of the Hessian Q. This is in contrast to the convergence rate which only depends on the extreme eigenvalues of the Hessian.

Next, we use Theorem 1 to determine the maximum of J among all quadratic functions $f \in \mathcal{F}_m^L$, which yields q defined in (4).

Proposition 1: Among all quadratic functions $f \in \mathcal{F}_m^L$, the largest steady-state second-order moment of $x^t \in \mathbb{R}^n$ for gradient descent and Nesterov's accelerated method with parameters provided in Table I is given by

$$\begin{aligned} q_{\mathrm{gd}} \,&=\, \frac{n\,\kappa^2}{2\,\kappa - 1} &&=\, n\,\Theta(\kappa) \\ q_{\mathrm{na}} \,&=\, \frac{n\,\kappa^2\,\left(2\,\kappa - 2\sqrt{\kappa} + 1\right)}{\left(2\sqrt{\kappa} - 1\right)^3} &&=\, n\,\Theta(\kappa\sqrt{\kappa}) \end{aligned}$$

where $\kappa = L/m$ is the condition number associated with $\mathcal{F}_m^L$.

Proof: For both gradient descent and Nesterov's accelerated algorithm with parameters provided in Table I, it is easy to show that the function $\hat{J}(\lambda)$ attains its maximum at $\lambda = m$. This implies that, among all quadratic functions in $\mathcal{F}_m^L$, the steady-state second-order moment J is maximized when all n eigenvalues of the matrix Q are equal to m. The result follows by substituting m for each λ_i in Theorem 1 and using the parameters defined in Table I. ■

Remark 1: The condition number $\kappa = L/m$ in Proposition 1 is defined with respect to the set $\mathcal{F}_m^L$ and not with respect to the elements of this set. In particular, for quadratic functions in $\mathcal{F}_m^L$, this quantity should not be confused with the condition number of the Hessian. For example, while the function $f(x) = \frac{1}{2}\, x^T x$ belongs to $\mathcal{F}_m^L$ for any $L \geq 1$ and $m \leq 1$, its Hessian is $\nabla^2 f = I$ and its condition number is 1. In general, for any quadratic function $f \in \mathcal{F}_m^L$ the condition number of the Hessian matrix $\nabla^2 f$ is smaller than or equal to the condition number of the set $\mathcal{F}_m^L$.

Remark 2: If we were to impose the additional requirement that the condition number of the Hessian is equal to κ for quadratic $f \in \mathcal{F}_m^L$, then J would be maximized by letting $\lambda_{\max}(\nabla^2 f) = L$ and the rest of the eigenvalues equal to m. This yields

$$\begin{aligned} q_{\mathrm{gd}} \,&=\, \frac{(n-1)\kappa^2}{2\kappa - 1} \,+\, 1 \\ q_{\mathrm{na}} \,&=\, \frac{(n-1)\left(\kappa^2\,(2\,\kappa - 2\sqrt{\kappa} + 1)\right)}{\left(2\sqrt{\kappa} - 1\right)^3} \,+\, 1 \end{aligned}$$

and all trends obtained in Proposition 1 are preserved.

IV. Upper bounds on the second-order moments

We now utilize results from control theory to derive meaningful upper bounds on the steady-state second-order moment J of noisy algorithm (2). For strongly convex smooth objective functions, the gradient mapping satisfies certain quadratic inequality constraints. We first present a result based on quadratic Lyapunov functions that exploits this property of the gradient and formulates upper bounds on J as solutions to a semidefinite programing problem. As we demonstrate, this approach yields tight upper bounds for gradient descent. However, for large condition numbers ($\kappa > 100$) this method does not provide any upper-bounds on J for Nesterov's method. Inspired by [30], we modify this semidefinite program to search over an enlarged set of Lyapunov functions obtained by adding the objective function to standard quadratic terms. We then employ this modified semidefinite program to derive tight upper bounds on J for Nesterov's accelerated method as well.

For any function $f \in \mathcal{F}_m^L$, the mapping $\Delta\colon \mathbb{R}^n \to \mathbb{R}^n$

$$\Delta(y^t) \; := \; \nabla f(y^t) \, - \, m y^t$$

satisfies the quadratic inequality [29, Lemma 6]

$$\begin{bmatrix} y \, - \, y_0 \\ \Delta(y) \, - \, \Delta(y_0) \end{bmatrix}^T \Pi \begin{bmatrix} y \, - \, y_0 \\ \Delta(y) \, - \, \Delta(y_0) \end{bmatrix} \; \geq \; 0 \qquad (6)$$

for all y, $y_0 \in \mathbb{R}^n$, where the matrix Π is given by

$$\Pi \; := \; \begin{bmatrix} 0 & (L \, - \, m)I \\ (L \, - \, m)I & -2I \end{bmatrix}. \qquad (7)$$

Consider a state-space model

$$\begin{aligned} \psi^{t+1} \; &= \; A\psi^t \, + \, B_w w^t \, + \, B_u u^t \\ \begin{bmatrix} z^t \\ y^t \end{bmatrix} \; &= \; \begin{bmatrix} C_z \\ C_y \end{bmatrix} \psi^t, \qquad u^t \; = \; \Delta(y^t) \end{aligned} \qquad (8a)$$

that contains a feedback interconnection of linear and nonlinear components. Algorithm (2) can be brought into the state-space form (8a) by selecting

$$\psi^t \; := \; \begin{bmatrix} x^t \\ x^{t+1} \end{bmatrix}, \; z^t \; := \; x^t, \; y^t \; := \; -\beta x^t + (1+\beta) x^{t+1}$$

and defining the corresponding matrices as

$$\begin{aligned} A \; &= \; \begin{bmatrix} 0 & I \\ -\beta(1-\alpha m)I & (1+\beta)(1-\alpha m)I \end{bmatrix} \\ B_w \; &= \; \begin{bmatrix} 0 \\ I \end{bmatrix}, \; B_u \; = \; \begin{bmatrix} 0 \\ -\alpha I \end{bmatrix} \\ C_z \; &= \; \begin{bmatrix} I & 0 \end{bmatrix}, \; C_y \; = \; \begin{bmatrix} -\beta I & (1+\beta)I \end{bmatrix}. \end{aligned} \qquad (8b)$$

Since gradient descent can be obtained from (2) by setting $\beta = 0$, in the absence of acceleration we can alternatively use $\psi^t = z^t = y^t := x^t$ with the corresponding matrices

$$A = (1 \, - \, \alpha m)I, \; B_w = C_z = C_y = I, \; B_u = -\alpha I. \quad (8c)$$

In what follows, we demonstrate how property (6) of the mapping Δ allows us to exploit results from control theory to obtain upper bounds on J when system (8a) is driven by the white noise input w^t with zero mean and identity covariance.

Lemma 1 employs a quadratic Lyapunov function $V(\psi) = \psi^T X \psi$ and provides an upper bound on the steady-state second-order moment of z^t in system (8a) that is characterized by a solution to an LMI. In Section IV-A, we show that this characterization provides a tight upper bound for the gradient descent method with the stepsize $\alpha = 1/L$.

Lemma 1: Let the nonlinear function $u = \Delta(y)$ satisfy the quadratic inequality

$$\begin{bmatrix} y \\ u \end{bmatrix}^T \Pi \begin{bmatrix} y \\ u \end{bmatrix} \; \geq \; 0 \qquad (9)$$

for some matrix Π, let X be a positive semidefinite matrix, and let λ be a positive scalar such that system (8a) satisfies

$$\begin{aligned} &\begin{bmatrix} A^T X A - X + C_z^T C_z & A^T X B_u \\ B_u^T X A & B_u^T X B_u \end{bmatrix} \; + \\ &\lambda \begin{bmatrix} C_y^T & 0 \\ 0 & I \end{bmatrix} \Pi \begin{bmatrix} C_y & 0 \\ 0 & I \end{bmatrix} \; \preceq \; 0. \end{aligned} \qquad (10)$$

Then the steady-state second-order moment J of the output z^t is bounded by

$$J \; \leq \; \mathrm{trace}\,(B_w^T X B_w).$$

The proof of Lemma 1 is omitted due to space limitations. For Nesterov's accelerated method with parameters provided in Table I, computational experiments show that Lemma 1 does not yield sensible upper bounds for $J_{\mathrm{na}}^\star$ as LMI (10) becomes infeasible for large values of κ. This observation is consistent with the results of [29] where it was suggested that, apart from (6), additional quadratic inequalities should be used to further tighten the constraints on the gradient ∇f and reduce conservativeness.

In Lemma 2, we build on the results of [30] and present an alternative LMI that is obtained using a Lyapunov function of the form $V(\psi) = \psi^T X \psi + f(x) - f(x^\star)$ where X is a positive semidefinite matrix and f is the objective function in (1). The resulting approach allows us to establish an analytical upper bound on $J_{\mathrm{na}}^\star$ for Nesterov's accelerated method in Section IV-B. The proof of Lemma 2 is omitted due to space limitations and it will be reported elsewhere.

Lemma 2: Let the matrix $M(m, L, \alpha, \beta)$ be defined as

$$M \; := \; N_1^T \begin{bmatrix} L I & I \\ I & 0 \end{bmatrix} N_1 \, + \, N_2^T \begin{bmatrix} -m I & I \\ I & 0 \end{bmatrix} N_2$$

where

$$\begin{aligned} N_1 \; &:= \; \begin{bmatrix} \alpha m \beta I & -\alpha m (1+\beta) I & -\alpha I \\ -m\beta I & m(1+\beta) I & I \end{bmatrix} \\ N_2 \; &:= \; \begin{bmatrix} -\beta I & \beta I & 0 \\ -m\beta I & m(1+\beta) I & I \end{bmatrix}. \end{aligned}$$

Consider state-space model (8a)-(8b) of noisy algorithm (2) and let Π be given by (7). Then, for any positive semidefinite matrix X and scalars $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$ such that

$$\begin{aligned} &\begin{bmatrix} A^T X A - X + C_z^T C_z & A^T X B_u \\ B_u^T X A & B_u^T X B_u \end{bmatrix} \; + \\ &\lambda_1 \begin{bmatrix} C_y^T & 0 \\ 0 & I \end{bmatrix} \Pi \begin{bmatrix} C_y & 0 \\ 0 & I \end{bmatrix} \; + \; \lambda_2 M \; \preceq \; 0 \end{aligned} \qquad (11)$$

the steady-state second-order moment J of the output z^t is bounded by

$$J \; \leq \; n L \lambda_2 \; + \; \mathrm{trace}\,(B_w^T X B_w). \qquad (12)$$

Remark 3: Since LMI (11) simplifies to LMI (10) by setting $\lambda_2 = 0$, Lemma 2 is a relaxed version of Lemma 1. This modification is the key which allows us to obtain a tight upper bound on $J_{\mathrm{na}}^\star$ in Section IV-B.

Lemma 1 and Lemma 2 are used in Sections IV-A and IV-B to establish upper bounds on $J_{\mathrm{gd}}^\star$ and $J_{\mathrm{na}}^\star$, respectively. These can be combined with the lower bounds of Proposition 1 to prove the main result of the paper.

Theorem 2: For gradient descent and Nesterov's accelerated method with parameters provided in Table I, the supremum of the steady-state second-order moment of the error $x^k - x^\star \in \mathbb{R}^n$ over all functions $f \in \mathcal{F}_m^L$ satisfies

$$J_{\rm gd}^\star \;=\; q_{\rm gd}$$
$$q_{\rm na} \;\leq J_{\rm na}^\star \;\leq\; 4.08\, q_{\rm na}$$

where

$$\begin{aligned} q_{\rm gd} \;&=\; \frac{n\,\kappa^2}{2\,\kappa - 1} &&=\; n\,\Theta(\kappa) \\ q_{\rm na} \;&=\; \frac{n\,\kappa^2\;(2\,\kappa - 2\,\sqrt{\kappa} + 1)}{(2\,\sqrt{\kappa} - 1)^3} &&=\; n\,\Theta(\kappa\sqrt{\kappa}) \end{aligned}$$

and $\kappa = L/m$ is the condition number associated with $\mathcal{F}_m^L$.

Proof: The lower bounds on $J_{\rm gd}^\star$ and $J_{\rm na}^\star$ are established in Proposition 1 and the upper bounds are established in Propositions 2 and 3, respectively. ■

Theorem 2 shows that for the class of strongly convex functions with condition number κ, gradient descent outperforms Nesterov's accelerated method in terms of the largest noise amplification by a factor of $\sqrt{\kappa}$. This uncovers the fundamental performance limitation of Nesterov's accelerated method when the gradient evaluation is subject to additive stochastic uncertainties.

A. Gradient descent

If $y_0 = 0$ in (6) is the minimizer of the objective function $f \in \mathcal{F}_m^L$, then condition (9) in Lemma 1 holds for noisy algorithm (2) with Π defined in (7). Thus, we can use Lemma 1 to obtain an upper bound on $J^\star$ by solving LMI (10). This leads to the following upper bound for gradient descent.

Proposition 2: For gradient descent with $\alpha = 1/L$, the steady-state second-order moment of x^t is bounded by

$$J_{\rm gd} \;\leq\; \frac{n\,\kappa^2}{2\,\kappa \,-\, 1}$$

for all $f \in \mathcal{F}_m^L$ where $\kappa := L/m$ is the condition number associated with the set $\mathcal{F}_m^L$.

Proof: Let $f \in \mathcal{F}_m^L$. To obtain the best upper bound on J using Lemma 1, we minimize $\mathrm{trace}\,(B_w^T X B_w)$ subject to LMI (10), $X \succeq 0$, and $\lambda \geq 0$. For gradient descent, if we use representation (8c), then the negative definiteness of the $(1,1)$-block of LMI (10) implies that

$$X \;\succeq\; \frac{1}{\alpha\,m(2\,-\,\alpha\,m)}\,I \;=\; \frac{\kappa^2}{2\,\kappa - 1}\,I. \tag{13}$$

It is straightforward to show that the pair

$$X \;=\; \frac{\kappa^2}{2\kappa - 1}\,I,\;\; \lambda \;=\; \frac{1 - \alpha\,m}{m(2 - \alpha\,m)(L - m)} \tag{14}$$

is feasible as the left-hand-side of LMI (10) becomes

$$\begin{bmatrix} 0 & 0 \\ 0 & \frac{-1}{m^2(2\kappa\,-\,1)}\,I \end{bmatrix} \;\preceq\; 0.$$

Thus, X and λ given by (14) provide a solution to LMI (10). This demonstrates that inequality (13) is tight and that it provides the best achievable upper bound

$$J_{\rm gd}^\star \;\leq\; \mathrm{trace}\,(B_w^T\,X\,B_w) \;=\; n\,\kappa^2/(2\,\kappa\,-\,1)$$

which completes the proof. ■

Remark 4: Since the upper bound established in Proposition 2 holds for any $f \in \mathcal{F}_m^L$, it serves as an upper bound for the supremum $J^\star$ as well. This upper bound is equal to the lower bound obtained using the restriction to quadratic objective functions established in Propositions 1, i.e.,

$$q_{\rm gd} \;=\; \frac{n\,\kappa^2}{2\,\kappa\,-\,1}.$$

Therefore, we obtain the exact expression

$$J_{\rm gd}^\star \;=\; q_{\rm gd} \;=\; \frac{n\,\kappa^2}{2\,\kappa - 1}$$

which implies that there exists a quadratic objective function for which the steady-state second-order moment of $x^t - x^\star$ for gradient descent with $\alpha = 1/L$ is the largest in $\mathcal{F}_m^L$.

B. Nesterov's accelerated method

The best upper bound on $J_{\rm na}^\star$ that can be obtained using Lemma 2 is given by the optimal objective value $p_{\rm na}$ of the semidefinite program

$$\begin{aligned} p_{\rm na} \;:=\; &\inf_{X,\,\lambda_1,\,\lambda_2}\; n\,L\,\lambda_2 \;+\; \mathrm{trace}\,(B_w^T\,X\,B_w) \\ &\text{subject to LMI (11) with parameters of Table I} \\ &\qquad X \;\succeq\; 0,\;\; \lambda_1 \;\geq\; 0,\;\; \lambda_2 \;\geq\; 0. \end{aligned} \tag{15}$$

We can evaluate this upper bound by numerically solving problem (15) for any values of m and L. We recall that $q_{\rm na}$ is obtained by restricting optimization over quadratic strongly convex functions. Our computational experiments suggest that the ratio between $p_{\rm na}$ and $q_{\rm na}$ is bounded by a constant, i.e., $p_{\rm na}/q_{\rm na} \leq 4.07$. Motivated by this observation, Proposition 3 establishes that $J_{\rm na}$ is bounded from above by a scaled version of the lower bound $q_{\rm na}$. The proof relies on finding a sub-optimal feasible point for problem (15) which we omit due to space limitations.

Proposition 3: For Nesterov's accelerated method with parameters provided in Table (I), the steady-state second-order moment of the optimization variable $x^t \in \mathbb{R}^n$ is upper bounded by

$$J_{\rm na} \;\leq\; 4.08\,q_{\rm na}$$

for all functions $f \in \mathcal{F}_m^L$, where $q_{\rm na}$ is given in Proposition 1.

Remark 5: Since the upper bound established in Proposition 2 holds for any $f \in \mathcal{F}_m^L$, it also serves as an upper bound for the supremum $J_{\rm na}^\star$. Now, if we compare the lower bound $q_{\rm na}$ on $J_{\rm na}^\star$ established in Proposition 1 (that is obtained by restricting to the set of quadratic objective functions) and the above upper bound we obtain,

$$q_{\rm na} \;\leq\; J_{\rm na}^\star \;\leq\; 4.08\,q_{\rm na}.$$

Thus, both bounds are tight up to the constant factor. This implies that there is a quadratic objective function for which the steady-state second-order moment of $x^t - x^\star$ for Nesterov's algorithm is not smaller than 4.08 times the largest steady-state second-order moment $J^\star_{\rm na}$ among all $f \in \mathcal{F}_m^L$.

Remark 6: For $x^t \in \mathbb{R}^n$ and matrices (8b), LMI (11) is of size $3n \times 3n$. However, if we impose the additional constraint that the matrix X has the same block structure as A,

$$X = \left[\begin{array}{cc} x_1 I & x_0 I \\ x_0 I & x_2 I \end{array} \right]$$

for some scalars x_1, x_2, and x_0, then using appropriate permutation matrices, we can simplify LMI (10) into an LMI of size 3×3. It can be shown that this additional constraint comes without loss of generality. In particular, the optimal objective value $p_{\rm na}$ of problem (15) does not change if we require X to have this structure; see [29, Section 4.2] for a discussion of this lossless dimensionality reduction for LMI constraints with similar structure.

V. Concluding remarks

We analyze the steady-state properties of gradient descent and Nesterov's accelerated methods perturbed by an additive white noise with zero mean and identity covariance. In particular, we consider the standard parameters for smooth strongly convex optimization problems with condition number κ and employ an LMI-based approach to establish explicit upper bounds on noise amplification that only depend on κ and the problem size n. For both algorithms, restriction to quadratic objective functions provides lower bounds and demonstrates that the upper bounds are tight up to a constant factor. We show that the upper bound for Nesterov's accelerated method is larger than the upper bound for the standard gradient descent by a factor of $\sqrt{\kappa}$. This uncovers the fundamental performance limitation of Nesterov's accelerated method when the gradient evaluation is subject to additive stochastic uncertainties.

References

[1] 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.

[2] 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.

[3] L. Bottou, F. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *SIAM Rev.*, vol. 60, no. 2, pp. 223–311, 2018.

[4] 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.

[5] 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.

[6] Y. Nesterov, "A method for solving the convex programming problem with convergence rate $O(1/k^2)$," in *Dokl. Akad. Nauk SSSR*, vol. 27, no. 2, 1983, pp. 543–547.

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

[8] 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.

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

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

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

[12] 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.

[13] B. Bamieh, M. R. Jovanović, P. Mitra, and S. Patterson, "Coherence in large-scale networks: dimension dependent limitations of local feedback," *IEEE Trans. Automat. Control*, vol. 57, no. 9, pp. 2235–2249, 2012.

[14] R. Ge, F. Huang, C. Jin, and Y. Yuan, "Escaping from saddle points – online stochastic gradient for tensor decomposition," in *Conference on Learning Theory*, 2015, pp. 797–842.

[15] C. Jin, R. Ge, P. Netrapalli, S. Kakade, and M. Jordan, "How to escape saddle points efficiently," in *Proceedings of the 34th International Conference on Machine Learning, PMLR 70*, 2017, pp. 1724–1732.

[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. Opt.*, vol. 19, no. 4, pp. 1574–1609, 2009.

[19] O. Devolder, F. Glineur, and Y. Nesterov, "First-order methods of smooth convex optimization with inexact oracle," *Math. Program.*, vol. 146, no. 1-2, pp. 37–75, 2014.

[20] F. R. Bach, "Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression." *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 595–627, 2014.

[21] A. Dieuleveut, N. Flammarion, and F. Bach, "Harder, better, faster, stronger convergence rates for least-squares regression," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 3520–3570, 2017.

[22] E. Moulines and F. R. Bach, "Non-asymptotic analysis of stochastic approximation algorithms for machine learning," in *Proc. Neural Information Processing (NIPS)*, 2011, pp. 451–459.

[23] N. Tripuraneni, N. Flammarion, F. Bach, and M. I. Jordan, "Averaging stochastic gradient descent on Riemannian manifolds," 2018, arXiv:1802.09128.

[24] M. Schmidt, N. L. Roux, and F. R. Bach, "Convergence rates of inexact proximal-gradient methods for convex optimization," in *Proc. Neural Information Processing (NIPS)*, 2011, pp. 1458–1466.

[25] M. Baes, "Estimate sequence methods: extensions and approximations," *IFOR Internal report, ETH, Zürich, Switzerland*, 2009.

[26] A. d'Aspremont, "Smooth optimization with approximate gradient," *SIAM J. Opt.*, vol. 19, no. 3, pp. 1171–1183, 2008.

[27] J.-F. Aujol and C. Dossal, "Stability of over-relaxations for the forward-backward algorithm, application to FISTA," *SIAM J. Opt.*, vol. 25, no. 4, pp. 2408–2433, 2015.

[28] 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.

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

[30] 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.

[31] B. Hu and L. Lessard, "Dissipativity theory for Nesterov's accelerated method," in *Proceedings of the 34th International Conference on Machine Learning, PMLR 70*, 2017, pp. 1549–1557.

[32] 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.

[33] N. K. Dhingra, S. Z. Khong, and M. R. Jovanović, "The proximal augmented Lagrangian method for nonsmooth composite optimization," *IEEE Trans. Automat. Control*, 2018, doi:10.1109/TAC.2018.2867589.