Robustness of Accelerated First-Order Algorithms for Strongly Convex Optimization Problems

Hesameddin Mohammadi, *Student Member, IEEE*, Meisam Razaviyayn, and Mihailo R. Jovanović, *Fellow, IEEE*

***Abstract*—We study the robustness of accelerated first-order algorithms to stochastic uncertainties in gradient evaluation. Specifically, for unconstrained, smooth, strongly convex optimization problems, we examine the mean-squared error in the optimization variable when the iterates are perturbed by additive white noise. This type of uncertainty may arise in situations where an approximation of the gradient is sought through measurements of a real system or in a distributed computation over a network. Even though the underlying dynamics of first-order algorithms for this class of problems are nonlinear, we establish upper bounds on the mean-squared deviation from the optimal solution that are tight up to constant factors. Our analysis quantifies fundamental tradeoffs between noise amplification and convergence rates obtained via *any* acceleration scheme similar to Nesterov's or heavy-ball methods. To gain additional analytical insight, for strongly convex quadratic problems, we explicitly evaluate the steady-state variance of the optimization variable in terms of the eigenvalues of the Hessian of the objective function. We demonstrate that the entire spectrum of the Hessian, rather than just the extreme eigenvalues, influences robustness of noisy algorithms. We specialize this result to the problem of distributed averaging over undirected networks and examine the role of network size and topology on the robustness of noisy accelerated algorithms.**

***Index Terms*—Accelerated first-order algorithms, consensus networks, control for optimization, convex optimization, integral quadratic constraints, linear matrix inequalities (LMIs), noise amplification, second-order moments, semidefinite programming.**

I. Introduction

FIRST-ORDER algorithms are well suited for solving a broad range of optimization problems that arise in statistics, signal and image processing, control, and machine learning [1]–[5]. Among these algorithms, accelerated methods enjoy the optimal rate of convergence, and they are popular because of their low per-iteration complexity. There is a large body of literature dedicated to the convergence analysis of these methods under different stepsize selection rules [2], [5]–[9]. In many applications, however, the exact value of the gradient is not fully available, e.g., when the objective function is obtained via costly simulations (e.g., tuning of hyperparameters in supervised/unsupervised learning [10]–[12] and model-free optimal control [13]–[16]), when evaluation of the objective function relies on noisy measurements (e.g., real-time and embedded applications), or when the noise is introduced via communication between different agents (e.g., distributed computation over networks). Another related application arises in the context of (batch) stochastic gradient, 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, additive noise may be introduced deliberately in the context of nonconvex optimization to help the iterates escape saddle points and improve generalization [17], [18].

In all above situations, first-order algorithms only have access to noisy estimates of the gradient. This observation has motivated the robustness analysis of first-order algorithms under different types of noisy/inexact gradient oracles [19]–[24]. For example, in a deterministic noise scenario, an upper bound on the error in iterates for accelerated proximal gradient methods was established in [25]. This study showed that both proximal gradient and its accelerated variant can maintain their convergence rates, provided that the noise is bounded and that it vanishes fast enough. Moreover, it has been shown that in the presence of random noise, with the proper diminishing stepsize, acceleration can be achieved for general convex problems. However, in this case, optimal rates are *sublinear* [26].

In the context of stochastic approximation, while early results suggest to use a stepsize that is inversely proportional to the iteration number [20], a more robust behavior can be obtained by combining larger stepsizes with averaging [21], [27]–[29]. Utility of these averaging schemes and their modifications for solving quadratic optimization and manifold problems has been examined thoroughly in recent years [30]–[32]. Moreover, several studies have suggested that accelerated first-order

Manuscript received May 26, 2019; revised February 19, 2020; accepted July 2, 2020. Date of publication July 9, 2020; date of current version May 27, 2021. This work was supported in part by the National Science Foundation under Award ECCS 1708906 and Award ECCS 1809833 and in part by the Air Force Office of Scientific Research under Award FA9550-16-1-0009. Recommended by Associate Editor D. Regruto. *(Corresponding author: Mihailo R. Jovanović.)*

Hesameddin Mohammadi and Mihailo R. Jovanović are with the Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA 90007 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 90007 USA (e-mail: razaviya@usc.edu).

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TAC.2020.3008297

algorithms are more susceptible to errors in the gradient compared to their nonaccelerated counterparts [22], [23], [25], [33]–[35].

One of the basic sources of error that arises in computing the gradient can be modeled by additive white stochastic noise. This source of error is typical for problems in which the gradient is being sought through measurements of a real system [36], and it has a rich history in analysis of stochastic dynamical systems and control theory [37]. Moreover, in many applications, including distributed computing over networks [38], [39], coordination in vehicular formations [40], and control of power systems [41], additive white noise is a convenient abstraction for the robustness analysis of distributed control strategies [39] and of first-order optimization algorithms [42], [43]. Motivated by this observation, in this article, we consider the scenario in which a white stochastic noise with zero mean and identity covariance is added to the iterates of standard first-order algorithms: gradient descent, Polyak's heavy-ball method, and Nesterov's accelerated algorithm. By confining our attention to smooth strongly convex problems, we provide a tight quantitative characterization for the mean-squared error of the optimization variable. Since this quantity provides a measure of how noise gets amplified by the dynamics resulting from optimization algorithms, we also refer to it as *noise* (or *variance*) *amplification.* We demonstrate that our quantitative characterization allows us to identify fundamental tradeoffs between the noise amplification and the rate of convergence obtained via acceleration.

This article is based on our recent conference papers [44], [45]. In a concurrent work [46], a similar approach was taken to analyze the robustness of gradient descent and Nesterov's accelerated method. Therein, it was shown that for a given convergence rate, one can select the algorithmic parameters such that the steady-state mean-squared error in the *objective value* of a Nesterov-like method becomes smaller than that of gradient descent. This is not surprising because gradient descent can be viewed as a special case of Nesterov's method with a zero momentum parameter. Using this argument, similar assertions have been made about the variance amplification of the *iterates*. This observation has been used to design an optimal multistage algorithm that does not require any information about the variance of the noise [47]. On the contrary, we demonstrate that there are fundamental differences between these two robustness measures, i.e., *objective values* and *iterates*, as the former does not capture the negative impact of acceleration in the presence of noise.

Focusing on the error in the iterates, we show that any choice of parameters for Nesterov's or heavy-ball methods that yields an accelerated convergence rate increases variance amplification relative to gradient descent. More precisely, *for the problem with the condition number κ, an algorithm with accelerated convergence rate of at least $1 - c/\sqrt{\kappa}$, where c is a positive constant, increases the variance amplification in the iterates by a factor of $\sqrt{\kappa}$.* The robustness problem was also studied in [48], where the authors show a similar behavior of Nesterov's method and gradient descent in an asymptotic regime, in which the stepsize goes to zero. In contrast, we focus on the nonasymptotic stepsize regime and establish fundamental differences between gradient descent and its accelerated variants in terms of noise amplification.

More recently, the problem of finding upper bounds on the variance amplification was cast as a semidefinite program [49]. This formulation provided numerical results that are consistent with our theoretical upper bounds in terms of the condition number. In [49], structured objective functions (e.g., diagonal Hessians) that arise in distributed optimization were also studied, and the problem of designing robust algorithms were formulated as a bilinear matrix inequality (which, in general, is not convex).

Contributions: The effect of imperfections on the performance and robustness of first-order algorithms has been studied in [23] and [31], but the influence of acceleration on stochastic gradient perturbations has not been precisely characterized. We employ control-theoretic tools suitable for analyzing stochastic dynamical systems to quantify such influence and identify fundamental tradeoffs between acceleration and noise amplification. The main contributions of this article are the following.

1) We start our analysis by examining strongly convex quadratic optimization problems for which we can explicitly characterize variance amplification of first-order algorithms and obtain analytical insight. In contrast to convergence rates, which solely depend on the extreme eigenvalues of the Hessian matrix, we demonstrate that the *variance amplification is influenced by the entire spectrum.*
2) We establish the relation between the noise amplification of accelerated algorithms and gradient descent for parameters that provide the optimal convergence rate for strongly convex quadratic problems. We also explain how the distribution of the eigenvalues of the Hessian influences these relations and provides examples to show that *acceleration can significantly increase amplification of noise.*
3) We address the problem of tuning the algorithmic parameters and demonstrate the existence of a fundamental tradeoff between the rate of convergence and noise amplification: for problems with condition number κ and bounded dimension n, we show that any choice of parameters in accelerated methods that yields the linear convergence rate of at least $1 - c/\sqrt{\kappa}$, where c is a positive constant, *increases noise amplification in the iterates relative to gradient descent* by a factor of at least $\sqrt{\kappa}$.
4) We extend our analysis from quadratic objective functions to general strongly convex problems. We borrow an approach based on linear matrix inequalities (LMIs) from control theory to establish upper bounds on the noise amplification of both gradient descent and Nesterov's accelerated algorithm. Furthermore, for any given condition number, we demonstrate that *these bounds are tight up to constant factors.*
5) We apply our results to distributed averaging over large-scale undirected networks. We examine the role of network size and topology on noise amplification and further illustrate the subtle influence of the entire spectrum of the Hessian matrix on the robustness of noisy optimization algorithms. In particular, *we identify a class of large-scale problems for which accelerated Nesterov's method achieves the same orderwise noise amplification* (in terms of condition number) *as gradient descent.*

Article structure: The rest of this article is organized as follows. In Section II, we formulate the problem and provide background material. In Section III, we explicitly evaluate the variance amplification (in terms of the algorithmic parameters and problem data) for strongly convex quadratic problems, derive lower and upper bounds, and provide a comparison between the accelerated methods and gradient descent. In Section IV, we extend our analysis to general strongly convex problems. In Section V, we establish fundamental tradeoffs between the rate of convergence and noise amplification. In Section VI, we apply our results to the problem of distributed averaging over noisy undirected networks. We highlight the subtle influence of the distribution of the eigenvalues of the Laplacian matrix on variance amplification and discuss the roles of network size and topology. Section VII concludes this article.

TABLE I

CONVENTIONAL VALUES OF PARAMETERS AND THE CORRESPONDING RATES FOR $f \in \mathcal{F}_m^L$, $\|x^t - x^\star\| \le c\,\rho^t\,\|x^0 - x^\star\|$, WHERE $\kappa := L/m$ AND $c > 0$ IS A CONSTANT [9, THEOREMS 2.1.15 AND 2.2.1]

Method	Parameters	Linear rate
Gradient	$\alpha = \frac{1}{L}$	$\rho = \sqrt{1 - \frac{2}{\kappa+1}}$
Nesterov	$\alpha = \frac{1}{L}$, $\beta = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$	$\rho = \sqrt{1 - \frac{1}{\sqrt{\kappa}}}$

The heavy-ball method does not offer acceleration guarantees for all $f \in \mathcal{F}_m^L$.

II. PRELIMINARIES AND BACKGROUND

In this article, we quantify the effect of stochastic uncertainties in gradient evaluation on the performance of first-order algorithms for unconstrained optimization problems

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

where $f\colon \mathbb{R}^n \to \mathbb{R}$ is strongly convex with Lipschitz continuous gradient ∇f. Specifically, we examine how gradient descent

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

Polyak's heavy-ball method

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

and Nesterov's accelerated method

$$\begin{aligned} x^{t+2} &= x^{t+1} + \beta(x^{t+1} - x^t) \\ &\quad - \alpha \nabla f\left(x^{t+1} + \beta(x^{t+1} - x^t)\right) + \sigma w^t \end{aligned} \tag{2c}$$

amplify the additive white stochastic noise w^t with zero mean and identity covariance matrix, i.e., $\mathbb{E}[w^t] = 0$ and $\mathbb{E}[w^t (w^\tau)^T] = I\,\delta(t - \tau)$. Here, t is the iteration index, x^t is the optimization variable, α is the stepsize, β is an extrapolation parameter used for acceleration, σ is the noise magnitude, δ is the Kronecker delta, and $\mathbb{E}$ is the expected value. When the only source of uncertainty is a noisy gradient, we set $\sigma = \alpha$ in (2).

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

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

In the absence of noise (i.e., for $\sigma = 0$), for $f \in \mathcal{F}_m^L$, the parameters α and β can be selected such that gradient descent and Nesterov's accelerated method converge to the global minimum $x^\star$ of (1) with a linear rate $\rho < 1$, i.e.,

$$\|x^t - x^\star\| \le c\,\rho^t \|x^0 - x^\star\|$$

for all t and some $c > 0$. Table I provides the conventional values of these parameters and the corresponding guaranteed convergence rates [9]. Nesterov's method with the parameters provided in Table I enjoys the convergence rate $\rho_{\rm na} = \sqrt{1 - 1/\sqrt{\kappa}} \le 1 - 1/(2\sqrt{\kappa})$, where $\kappa := L/m$ is the condition number associated with $\mathcal{F}_m^L$. This rate is *orderwise optimal* in the sense that no first-order algorithm can optimize all $f \in \mathcal{F}_m^L$ with the rate $\rho_{\rm lb} = (\sqrt{\kappa} - 1)/(\sqrt{\kappa} + 1)$ [9, Th. 2.1.13]. Note that $1 - \rho_{\rm lb} = O(1/\sqrt{\kappa})$ and $1 - \rho_{\rm na} = \Omega(1/\sqrt{\kappa})$. In contrast to Nesterov's method, the heavy-ball method does not offer any acceleration guarantees for all $f \in \mathcal{F}_m^L$. However, for strongly convex quadratic f, parameters can be selected to guarantee linear convergence of the heavy-ball method with a rate that outperforms the one achieved by Nesterov's method [50] (see Table II).

To provide a quantitative characterization for the robustness of algorithms (2) to the noise w^t, we examine the performance measure

$$J := \limsup_{t \to \infty} \frac{1}{t} \sum_{k=0}^{t} \mathbb{E}\left(\|x^k - x^\star\|^2\right). \tag{3}$$

For quadratic objective functions, algorithms (2) are linear dynamical systems. In this case, J quantifies the steady-state variance amplification, and it can be computed from the solution of the algebraic Lyapunov equation (see Section III). For general strongly convex problems, there is no explicit characterization for J, but techniques from control theory can be utilized to compute an upper bound (see Section IV).

Notation: We write $g = \Omega(h)$ (or, equivalently, $h = O(g)$) to denote the existence of positive constants c_i such that, for any $x > c_2$, the functions g and h: $\mathbb{R} \to \mathbb{R}$ satisfy $g(x) \ge c_1 h(x)$. We write $g = \Theta(h)$, or more informally $g \approx h$, if both $g = \Omega(h)$ and $g = O(h)$.

III. STRONGLY CONVEX QUADRATIC PROBLEMS

Consider a strongly convex quadratic objective function

$$f(x) = \tfrac{1}{2} x^T Q x - q^T x \tag{4}$$

where Q is a symmetric positive-definite matrix and q is a vector. Let $f \in \mathcal{F}_m^L$ and let the eigenvalues λ_i of Q satisfy

$$L \ = \lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_n = m \ > 0.$$

In the absence of noise, the constant values of parameters α and β provided in Table II yield linear convergence (with optimal decay rates) to the globally optimal point $x^\star = Q^{-1} q$ for all three algorithms [50]. In the presence of additive white noise w^t, we derive analytical expressions for the variance amplification J of algorithms (2) and demonstrate that J depends not only on the algorithmic parameters α and β, but also on all eigenvalues of the Hessian matrix Q. This should be compared and contrasted to the optimal rate of linear convergence, which only depends on $\kappa := L/m$, i.e., the ratio of the largest and smallest eigenvalues of Q.

For constant α and β, algorithms (2) can be described by a linear time-invariant (LTI) first-order recursion

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

TABLE II

OPTIMAL PARAMETERS AND THE CORRESPONDING CONVERGENCE RATES FOR A STRONGLY CONVEX QUADRATIC OBJECTIVE FUNCTION $f \in \mathcal{F}_m^L$ WITH $\lambda_{\max}(\nabla^2 f) = L$ AND $\lambda_{\min}(\nabla^2 f) = m$, AND $\kappa := L/m$ [50, PROP. 1]

Method	Optimal parameters	Rate of linear convergence
Gradient	$\alpha = \frac{2}{L+m}$	$\rho = \frac{\kappa-1}{\kappa+1}$
Nesterov	$\alpha = \frac{4}{3L+m}$, $\beta = \frac{\sqrt{3\kappa+1}-2}{\sqrt{3\kappa+1}+2}$	$\rho = \frac{\sqrt{3\kappa+1}-2}{\sqrt{3\kappa+1}}$
Heavy-ball	$\alpha = \frac{4}{(\sqrt{L}+\sqrt{m})^2}$, $\beta = \frac{(\sqrt{\kappa}-1)^2}{(\sqrt{\kappa}+1)^2}$	$\rho = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$

where ψ^t is the state, $z^t := x^t - x^\star$ is the performance output, and w^t is a white stochastic input. In particular, choosing $\psi^t := x^t - x^\star$ for gradient descent and $\psi^t := [(x^t - x^\star)^T (x^{t+1} - x^\star)^T]^T$ for accelerated algorithms yields state-space model (5) with

$$A = I - \alpha Q\,, B = C \;=\; I$$

for gradient descent and

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

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

for the heavy-ball and Nesterov's methods, respectively, with

$$B^T = [\,0 \quad I\,]\,, C = [\,I \quad 0\,]\,.$$

Since w^t is zero mean, we have $\mathbb{E}(\psi^{t+1}) = A\,\mathbb{E}(\psi^t)$. Thus, $\mathbb{E}(\psi^t) = A^t\,\mathbb{E}(\psi^0)$ and, for any stabilizing parameters α and β, $\lim_{t\to\infty} \mathbb{E}(\psi^t) = 0$, with the same linear rate as in the absence of noise. Furthermore, it is well known that the covariance matrix $P^t := \mathbb{E}(\psi^t(\psi^t)^T)$ of the state vector satisfies the linear recursion

$$P^{t+1} = AP^tA^T + \sigma^2 BB^T \tag{6a}$$

and that its steady-state limit

$$P := \lim_{t\to\infty} \mathbb{E}\left(\psi^t(\psi^t)^T\right) \tag{6b}$$

is the unique solution to the algebraic Lyapunov equation [37]

$$P = APA^T + \sigma^2 BB^T. \tag{6c}$$

For stable LTI systems, performance measure (3) simplifies to the steady-state variance of the error in the optimization variable $z^t := x^t - x^\star$

$$J = \lim_{t\to\infty} \frac{1}{t} \sum_{k=0}^{t} \mathbb{E}\left(\|z^k\|^2\right) = \lim_{t\to\infty} \mathbb{E}\left(\|z^t\|^2\right) \tag{6d}$$

and it can be computed using either of the following two equivalent expressions:

$$J = \lim_{t\to\infty} \frac{1}{t} \sum_{k=0}^{t} \text{trace}\left(Z^k\right) = \text{trace}(Z) \tag{6e}$$

where $Z = CPC^T$ is the steady-state limit of the output covariance matrix $Z^t := \mathbb{E}(z^t(z^t)^T) = CP^tC^T$.

We next provide analytical solution P to (6c) that depends on the parameters α and β as well as on the spectrum of the Hessian matrix Q. This allows us to explicitly characterize the variance amplification J and quantify the impact of additive white noise on the performance of first-order optimization algorithms.

A. Influence of the Eigenvalues of the Hessian Matrix

We use the modal decomposition of the symmetric matrix $Q = V\Lambda V^T$ to bring A, B, and C in (5) into a block diagonal form, $\hat{A} = \text{diag}\,(\hat{A}_i)$,$\hat{B} = \text{diag}\,(\hat{B}_i)$,$\hat{C} = \text{diag}\,(\hat{C}_i)$, with $i = 1, \ldots, n$. Here, $\Lambda = \text{diag}\,(\lambda_i)$ is the diagonal matrix of the eigenvalues and V is the orthogonal matrix of the eigenvectors of Q. More specifically, the unitary coordinate transformation

$$\hat{x}^t \;:=\; V^Tx^t, \hat{x}^\star \;:=\; V^Tx^\star, \hat{w}^t \;:=\; V^Tw^t \tag{7}$$

brings the state-space model of gradient descent into a diagonal form with

$$\hat{\psi}_i^t = \hat{x}_i^t \,-\, \hat{x}_i^\star, \hat{A}_i = 1 - \alpha\lambda_i, \hat{B}_i = \hat{C}_i = 1. \tag{8a}$$

Similarly, for Polyak's heavy-ball and Nesterov's accelerated methods, change of coordinates (7) in conjunction with a permutation of variables, $\hat{\psi}_i^t = [\hat{x}_i^t - \hat{x}_i^\star \;\, \hat{x}_i^{t+1} - \hat{x}_i^\star]^T$, respectively, yields

$$\hat{A}_i \;=\; \begin{bmatrix} 0 & 1 \\ -\beta & 1 + \beta - \alpha\lambda_i \end{bmatrix} \tag{8b}$$

$$\hat{A}_i \;=\; \begin{bmatrix} 0 & 1 \\ -\beta(1 - \alpha\lambda_i) & (1+\beta)(1 - \alpha\lambda_i) \end{bmatrix} \tag{8c}$$

$$\hat{B}_i^T \;= [\,0 \quad 1\,]\,, \hat{C}_i = [\,1 \quad 0\,]\,. \tag{8d}$$

This block diagonal structure allows us to explicitly solve Lyapunov equation (6c) for P and derive an analytical expression for J in terms of the eigenvalues λ_i of the Hessian matrix Q and the algorithmic parameters α and β. Namely, under coordinate transformation (7) and a suitable permutation of variables, (6c) can be brought into an equivalent set of equations

$$\hat{P}_i = \hat{A}_i\hat{P}_i\hat{A}_i^T + \sigma^2\hat{B}_i\hat{B}_i^T, \quad i = 1, \ldots, n \tag{9}$$

where $\hat{P}_i$ is a scalar for the gradient descent method and a 2×2 matrix for the accelerated algorithms. In Theorem 1, we use the solution to these decoupled Lyapunov equations to express the variance amplification as

$$J = \sum_{i=1}^{n} \hat{J}(\lambda_i) \;:=\; \sum_{i=1}^{n} \text{trace}\,(\hat{C}_i\hat{P}_i\hat{C}_i^T)$$

where $\hat{J}(\lambda_i)$ determines the contribution of the eigenvalue λ_i of the matrix Q to the variance amplification. In what follows, we use subscripts gd, hb, and na (e.g., J_{gd}, J_{hb}, and J_{na}) to denote quantities that correspond to gradient descent (2a), heavy-ball method (2b), and Nesterov's accelerated method (2c).

Theorem 1: For strongly convex quadratic problems, the variance amplification of noisy first-order algorithms (2) with any constant stabilizing parameters α and β is determined by $J = \sum_{i=1}^{n} \hat{J}(\lambda_i)$, where λ_i is the ith eigenvalue of $Q = Q^T \succ 0$ and the modal contribution to the variance amplification $\hat{J}(\lambda)$ is given by

$$\begin{aligned}\hat{J}_{\text{gd}}(\lambda) &= \frac{\sigma^2}{\alpha\lambda\,(2-\alpha\lambda)}\\ \hat{J}_{\text{hb}}(\lambda) &= \frac{\sigma^2(1+\beta)}{\alpha\lambda\,(1-\beta)\,(2(1+\beta)-\alpha\lambda)}\\ \hat{J}_{\text{na}}(\lambda) &= \frac{\sigma^2(1+\beta(1-\alpha\lambda))}{\alpha\lambda\,(1-\beta(1-\alpha\lambda))\,(2(1+\beta)-(2\beta+1)\alpha\lambda)}.\end{aligned}$$

Proof: See Appendix A. ∎

For strongly convex quadratic problems, Theorem 1 provides *exact expressions* for variance amplification of the first-order algorithms. These expressions not only quantify the dependence of J on the algorithmic parameters α and β and the impact of the largest and smallest eigenvalues, but also capture the effect of all other eigenvalues of the Hessian matrix Q. We also observe that the variance amplification J is proportional to σ^2. Apart from Section V, where we examine the role of parameters α and β on acceleration and variance amplification, without loss of generality, we choose $\sigma = 1$ in the rest of this article.

Remark 1: The performance measure J in (6d) quantifies the steady-state variance of the iterates of first-order algorithms. Robustness of noisy algorithms can also be evaluated using alternative performance measures, e.g., the mean value of the error in the objective function [46]

$$J' \;=\; \lim_{t\to\infty} \mathbb{E}\left((x^t - x^\star)^T Q (x^t - x^\star)\right). \tag{10}$$

This measure of variance amplification can be characterized using our approach by defining $C = Q^{1/2}$ for gradient descent and $C = [Q^{1/2}\;\; 0]$ for accelerated algorithms in the state-space model (5). Furthermore, repeating the above procedure for the modified performance output z^t yields $J' = \sum_{i=1}^{n} \lambda_i \hat{J}(\lambda_i)$, where the respective expressions for $\hat{J}(\lambda_i)$ are given in Theorem 1.

B. Comparison for the Parameters That Optimize the Convergence Rate

We next examine the robustness of first-order algorithms applied to strongly convex quadratic problems for the parameters that optimize the linear convergence rate (see Table II). For these parameters, the eigenvalues of the matrix A are inside the open unit disk, implying exponential stability of system (5). We first use the expressions presented in Theorem 1 to compare the variance amplification of the heavy-ball method to gradient descent.

Theorem 2: Let the strongly convex quadratic objective function f in (4) satisfy $\lambda_{\max}(Q) = L$, $\lambda_{\min}(Q) = m > 0$, and let $\kappa := L/m$ be the condition number. For the optimal parameters provided in Table II, the ratio between the variance amplification of the heavy-ball method and gradient descent with equal values of σ is given by

$$\frac{J_{\text{hb}}}{J_{\text{gd}}} = \frac{(\sqrt{\kappa}+1)^4}{8\sqrt{\kappa}(\kappa+1)}. \tag{11}$$

Proof: For the parameters provided in Table II, we have $\alpha_{\text{hb}} = (1+\beta)\alpha_{\text{gd}}$, where $\beta = (\sqrt{\kappa}-1)^2/(\sqrt{\kappa}+1)^2$ is the momentum parameter for the heavy-ball method. It is now straightforward to show that the modal contributions $\hat{J}_{\text{hb}}$ and $\hat{J}_{\text{gd}}$ to the variance amplification of the iterates given in Theorem 1 satisfy

$$\frac{\hat{J}_{\text{hb}}(\lambda)}{\hat{J}_{\text{gd}}(\lambda)} = \frac{1}{1-\beta^2} = \frac{(\sqrt{\kappa}+1)^4}{8\sqrt{\kappa}(\kappa+1)} \quad \forall \lambda \in [m, L]. \tag{12}$$

Thus, *the ratio $\hat{J}_{\text{hb}}(\lambda)/\hat{J}_{\text{gd}}(\lambda)$ does not depend on λ and is only a function of the condition number κ.* Substitution of (12) into $J = \sum_i \hat{J}(\lambda_i)$ yields relation (11). ∎

Theorem 2 establishes the linear relation between the variance amplification of the heavy-ball algorithm J_{hb} and the gradient descent J_{gd}. We observe that the ratio $J_{\text{hb}}/J_{\text{gd}}$ *only* depends on the condition number κ and that *acceleration increases variance amplification*: for $\kappa \gg 1$, J_{hb} is larger than J_{gd} by a factor of $\sqrt{\kappa}$. We next study the ratio between the variance amplification of Nesterov's accelerated method and gradient descent. In contrast to the heavy-ball method, this ratio depends on the entire spectrum of the Hessian matrix Q. The following proposition, which examines the modal contributions $\hat{J}_{\text{na}}(\lambda)$ and $\hat{J}_{\text{gd}}(\lambda)$ of Nesterov's accelerated method and gradient descent, is the key technical result that allows us to establish the largest and smallest values that the ratio $J_{\text{na}}/J_{\text{gd}}$ can take for a given pair of extreme eigenvalues m and L of Q in Theorem 3.

Proposition 1: Let the strongly convex quadratic objective function f in (4) satisfy $\lambda_{\max}(Q) = L$, $\lambda_{\min}(Q) = m > 0$, and let $\kappa := L/m$ be the condition number. For the optimal parameters provided in Table II, the ratio $\hat{J}_{\text{na}}(\lambda)/\hat{J}_{\text{gd}}(\lambda)$ of modal contributions to variance amplification of Nesterov's method and gradient descent is a decreasing function of $\lambda \in [m, L]$. Furthermore, for $\sigma = 1$, the function $\hat{J}_{\text{gd}}(\lambda)$ satisfies

$$\begin{aligned}\max_{\lambda\in[m,L]} \hat{J}_{\text{gd}}(\lambda) &= \hat{J}_{\text{gd}}(m) = \hat{J}_{\text{gd}}(L) = \frac{(\kappa+1)^2}{4\kappa}\\ \min_{\lambda\in[m,L]} \hat{J}_{\text{gd}}(\lambda) &= \hat{J}_{\text{gd}}(1/\alpha) = 1\end{aligned} \tag{13a}$$

and the function $\hat{J}_{\text{na}}(\lambda)$ satisfies

$$\begin{aligned}\max_{\lambda\in[m,L]} \hat{J}_{\text{na}}(\lambda) &= \hat{J}_{\text{na}}(m) = \frac{\bar{\kappa}^2\left(\bar{\kappa} - 2\sqrt{\bar{\kappa}} + 2\right)}{32\left(\sqrt{\bar{\kappa}}-1\right)^3}\\ \min_{\lambda\in[m,L]} \hat{J}_{\text{na}}(\lambda) &= \hat{J}_{\text{na}}(1/\alpha) = 1\\ \hat{J}_{\text{na}}(L) &= \frac{9\bar{\kappa}^2\left(\bar{\kappa} + 2\sqrt{\bar{\kappa}} - 2\right)}{32\,(\bar{\kappa}-1)\left(\bar{\kappa} - \sqrt{\bar{\kappa}} + 1\right)\left(2\sqrt{\bar{\kappa}}-1\right)}\end{aligned} \tag{13b}$$

where $\bar{\kappa} := 3\kappa + 1$.

Proof: See Appendix A. ∎

For all three algorithms, Proposition 1 and Theorem 2 demonstrate that the modal contribution to the variance amplification of the iterates at the extreme eigenvalues of the Hessian matrix, m and L, only depends on the condition number $\kappa := L/m$. For gradient descent and the heavy-ball method, $\hat{J}$ achieves its largest value at m and L, i.e.,

$$\begin{aligned}&\max_{\lambda\in[m,L]} \hat{J}_{\rm gd}(\lambda) = \hat{J}_{\rm gd}(m) = \hat{J}_{\rm gd}(L) = \Theta(\kappa)\\ &\max_{\lambda\in[m,L]} \hat{J}_{\rm hb}(\lambda) = \hat{J}_{\rm hb}(m) = \hat{J}_{\rm hb}(L) = \Theta(\kappa\sqrt{\kappa}).\end{aligned} \tag{14a}$$

In contrast, for Nesterov's method, (13b) implies a gap of $\Theta(\kappa)$ between the boundary values

$$\max_{\lambda\in[m,L]} \hat{J}_{\rm na}(\lambda) = \hat{J}_{\rm na}(m) = \Theta(\kappa\sqrt{\kappa}), \hat{J}_{\rm na}(L) = \Theta(\sqrt{\kappa}). \tag{14b}$$

Remark 2: Theorem 1 provides explicit formulas for variance amplification of noisy algorithms (2) in terms of the eigenvalues λ_i of the Hessian matrix Q. Similarly, we can represent the variance amplification in terms of the eigenvalues $\hat{\lambda}_i$ of the dynamic matrices $\hat{A}_i$ in (8). For gradient descent, $\hat{\lambda}_i = 1 - \alpha\lambda_i$, and it is straightforward to verify that $J_{\rm gd}$ is determined by the sum of reciprocals of distances of these eigenvalues to the stability boundary, $J_{\rm gd} = \sum_{i=1}^{n} \sigma^2/(1-\hat{\lambda}_i^2)$. Similarly, for accelerated methods, we have,

$$J \;=\; \sum_{i=1}^{n} \frac{\sigma^2(1+\hat{\lambda}_i\hat{\lambda}_i')}{(1-\hat{\lambda}_i\hat{\lambda}_i')(1-\hat{\lambda}_i)(1-\hat{\lambda}_i')(1+\hat{\lambda}_i)(1+\hat{\lambda}_i')}$$

where $\hat{\lambda}_i$ and $\hat{\lambda}_i'$ are the eigenvalues of $\hat{A}_i$. For Nesterov's method with the parameters provided in Table II, the matrix $\hat{A}_n$, which corresponds to $\lambda_n = m$, admits a Jordan canonical form with repeated eigenvalues $\hat{\lambda}_n = \hat{\lambda}_n' = 1 - 2/\sqrt{3\kappa+1}$. In this case, $\hat{J}_{\rm na}(m) = \sigma^2(1+\hat{\lambda}_n^2)/(1-\hat{\lambda}_n^2)^3$, which should be compared and contrasted to the above expression for gradient descent. Furthermore, for both $\lambda_1 = L$ and $\lambda_n = m$, the matrices $\hat{A}_1$ and $\hat{A}_n$ for the heavy-ball method with the parameters provided in Table II have eigenvalues with algebraic multiplicity two and incomplete sets of eigenvectors.

We next establish the range of values that $J_{\rm na}/J_{\rm gd}$ can take.

Theorem 3: For the strongly convex quadratic objective function f in (4) with $x \in \mathbb{R}^n$, $\lambda_{\max}(Q) = L$, and $\lambda_{\min}(Q) = m > 0$, the ratio between the variance amplification of Nesterov's accelerated method and gradient descent, for the optimal parameters provided in Table II and equal values of σ, satisfies

$$\frac{J_{\rm na}}{J_{\rm gd}} \;\le\; \frac{\hat{J}_{\rm na}(L) + (n\,-\,1)\hat{J}_{\rm na}(m)}{\hat{J}_{\rm gd}(L) + (n\,-\,1)\hat{J}_{\rm gd}(m)} \tag{15a}$$

$$\frac{J_{\rm na}}{J_{\rm gd}} \;\ge\; \frac{\hat{J}_{\rm na}(m) + (n\,-\,1)\hat{J}_{\rm na}(L)}{\hat{J}_{\rm gd}(m) + (n\,-\,1)\hat{J}_{\rm gd}(L)}. \tag{15b}$$

Proof: See Appendix A. ■

Theorem 3 provides tight upper and lower bounds on the ratio between $J_{\rm na}$ and $J_{\rm gd}$ for strongly convex quadratic problems. As shown in Appendix A, the lower bound is achieved for a quadratic function in which the Hessian matrix Q has one eigenvalue at m and $n-1$ eigenvalues at L, and the upper bound is achieved when Q has one eigenvalue at L and the remaining ones at m. Theorem 3 in conjunction with Proposition 1 demonstrates that *for a fixed problem dimension n, $J_{\rm na}$ is larger than $J_{\rm gd}$ by a factor of $\sqrt{\kappa}$ for $\kappa \gg 1$.*

This tradeoff is further highlighted in Theorem 4, which provides tight bounds on the variance amplification of iterates in terms of the problem dimension n and the condition number κ for all three algorithms. To simplify the presentation, we first use the explicit expressions for $\hat{J}_{\rm na}(m)$ and $\hat{J}_{\rm na}(L)$ in Proposition 1 to obtain the following upper and lower bounds on $\hat{J}_{\rm na}(m)$ and $\hat{J}_{\rm na}(L)$ (see Appendix A)

$$\frac{(3\kappa+1)^{\frac{3}{2}}}{32} \;\le \hat{J}_{\rm na}(m) \;\le\; \frac{(3\kappa+1)^{\frac{3}{2}}}{8} \tag{16a}$$

$$\frac{9\sqrt{3\kappa+1}}{64} \;\le \hat{J}_{\rm na}(L) \;\le\; \frac{9\sqrt{3\kappa+1}}{8}. \tag{16b}$$

Theorem 4: For the strongly convex quadratic objective function f in (4) with $x \in \mathbb{R}^n$, $\lambda_{\max}(Q) = L$, $\lambda_{\min}(Q) = m > 0$, and $\kappa := L/m$, the variance amplification of the first-order optimization algorithms, with the parameters provided in Table II and $\sigma = 1$, is bounded by

$$\begin{aligned}\frac{(\kappa-1)^2}{2\kappa} + n \le J_{\rm gd} &\le \frac{n(\kappa+1)^2}{4\kappa}\\ J_{\rm hb} &\le \frac{n(\kappa+1)(\sqrt{\kappa}+1)^4}{32\kappa\sqrt{\kappa}}\\ J_{\rm hb} &\ge \frac{(\sqrt{\kappa}+1)^4}{8\sqrt{\kappa}(\kappa+1)}\left(\frac{(\kappa-1)^2}{2\kappa} + n\right)\\ J_{\rm na} &\le \frac{(n-1)(3\kappa+1)^{\frac{3}{2}}}{8} \;+\; \frac{9\sqrt{3\kappa+1}}{8}\\ J_{\rm na} &\ge \frac{(3\kappa+1)^{\frac{3}{2}}}{32} \;+\; \frac{9\sqrt{3\kappa+1}}{64} + n - 2.\end{aligned} \tag{17}$$

Proof: As shown in Proposition 1, the functions $\hat{J}(\lambda)$ for gradient descent and Nesterov's algorithm attain their largest and smallest values over the interval $[m, L]$ at $\lambda = m$ and $\lambda = 1/\alpha$, respectively. Thus, fixing the smallest and largest eigenvalues, the variance amplification J is maximized when the other $n-2$ eigenvalues are all equal to m and is minimized when they are all equal to $1/\alpha$. This combined with the explicit expressions for $\hat{J}_{\rm gd}(m)$, $\hat{J}_{\rm gd}(L)$, and $\hat{J}_{\rm gd}(1/\alpha)$ in (13a) leads to the tight upper and lower bounds for gradient descent. For the heavy-ball method, the bounds follow from Theorem 2, and for Nesterov's algorithm, the bounds follow from (16). ■

For problems with a fixed dimension n and a condition number $\kappa \gg n$, there is an $\Omega(\sqrt{\kappa})$ difference in both upper and lower bounds provided in Theorem 4 for the accelerated algorithms relative to gradient descent. Even though Theorem 4 considers only the values of α and β that optimize the convergence rate, in Section V, we demonstrate that this gap is fundamental in that it holds for any parameters that yield an accelerated convergence rate. It is worth noting that both the lower and upper bounds are influenced by the problem dimension n and the condition number κ. For large-scale problems, there may be a subtle relation between n and κ, and the established bounds may exhibit different scaling trends. In Section VI, we identify a class of quadratic optimization problems for which $J_{\rm na}$ scales in the same way as $J_{\rm gd}$ for $\kappa \gg 1$ and $n \gg 1$.

Before we elaborate further on these issues, we provide two illustrative examples that highlight the importance of the choice of the performance metric in the robustness analysis of noisy algorithms. It is worth noting that an $O(\kappa)$ upper bound for gradient descent and an $O(\kappa^2)$ upper bound for Nesterov's accelerated algorithm were established in [25]. Relative to this upper bound for Nesterov's method, the upper bound provided in

Theorem 4 is tighter by a factor of $\sqrt{\kappa}$. Theorem 4 also provides lower bounds, reveals the influence of the problem dimension n, and identifies constants that multiply the leading terms in the condition number κ. Moreover, in Section IV, we demonstrate that similar upper bounds can be obtained for general strongly convex objective functions with Lipschitz continuous gradients.

C. Examples

We next provide illustrative examples to 1) demonstrate the agreement of our theoretical predictions with the results of stochastic simulations and 2) contrast two natural performance measures, namely the variance of the iterates J in (6d) and the mean objective error J' in (10), for assessing robustness of noisy optimization algorithms.

Example 1: Let us consider the quadratic objective function in (4) with

$$Q = \begin{bmatrix} L & 0 \\ 0 & m \end{bmatrix}, q = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{18}$$

For all three algorithms, the performance measures J and J' are given by

$$\begin{aligned} J \;&=\; \hat{J}(m) + \hat{J}(L) \\ J' \;&=\; m\hat{J}(m) + L\hat{J}(L) \\ &=\; L\left(\tfrac{1}{\kappa}\hat{J}(m) + \hat{J}(L)\right) = m\left(\hat{J}(m) + \kappa\hat{J}(L)\right). \end{aligned}$$

As shown in (14), $\hat{J}(m)$ and $\hat{J}(L)$ only depend on the condition number κ, and the variance amplification of the iterates satisfies

$$J_{\rm gd} = \Theta(\kappa), J_{\rm hb} = \Theta(\kappa\sqrt{\kappa}), J_{\rm na} = \Theta(\kappa\sqrt{\kappa}). \tag{19a}$$

In contrast, J' also depends on m and L. In particular, it is easy to verify the following relations for two scenarios that yield $\kappa \gg 1$.

1) For $m \ll 1$ and $L = O(1)$, we have

$$J'_{\rm gd} = \Theta(\kappa), J'_{\rm hb} = \Theta(\kappa\sqrt{\kappa}), J'_{\rm na} = \Theta(\sqrt{\kappa}). \tag{19b}$$

2) For $L \gg 1$ and $m = O(1)$, we have

$$J'_{\rm gd} = \Theta(\kappa^2), J'_{\rm hb} = \Theta(\kappa^2\sqrt{\kappa}), J'_{\rm na} = \Theta(\kappa\sqrt{\kappa}). \tag{19c}$$

Relation (19a) reveals the detrimental impact of acceleration on the variance of the optimization variable. In contrast, (19b) and (19c) show that, relative to gradient descent, the heavy-ball method increases the mean error in the objective function, while Nesterov's method reduces it. Thus, if the mean value of the error in the objective function is to be used to assess performance of noisy algorithms, one can conclude that Nesterov's method significantly outperforms gradient descent both in terms of convergence rate and robustness to noise. However, this performance metric fails to capture large variance of the mode associated with the smallest eigenvalue of the matrix Q in Nesterov's algorithm. Theorem 2 and Proposition 1 show that the modal contributions to the variance amplification of the iterates for gradient descent and the heavy-ball method are balanced at m and L, i.e., $\hat{J}_{\rm gd}(m) = \hat{J}_{\rm gd}(L) = \Theta(\kappa)$ and $\hat{J}_{\rm hb}(m) = \hat{J}_{\rm hb}(L) = \Theta(\kappa\sqrt{\kappa})$. In contrast, for Nesterov's method, there is a $\Theta(\kappa)$ gap between $\hat{J}_{\rm na}(m) = \Theta(\kappa\sqrt{\kappa})$ and $\hat{J}_{\rm na}(L) = \Theta(\sqrt{\kappa})$. While the performance measure J' reveals a superior performance of Nesterov's algorithm at large condition numbers, it fails to capture the negative impact of acceleration on the variance of the optimization variable (see Fig. 1 for an illustration).

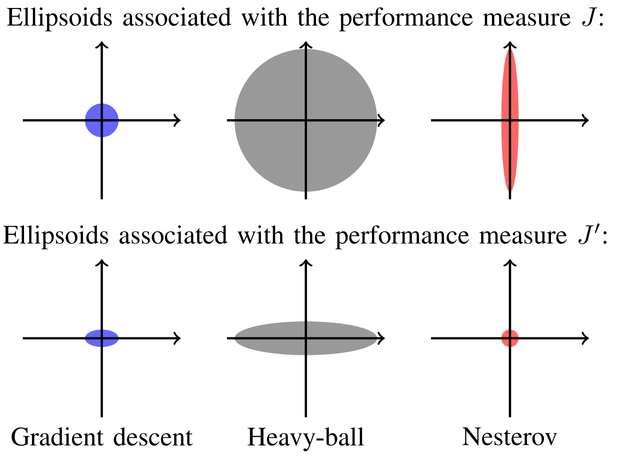
Ellipsoids associated with the performance measure J:
Ellipsoids associated with the performance measure J':
Gradient descent
Heavy-ball
Nesterov

Fig. 1. Ellipsoids $\{z | z^T Z^{-1} z \leq 1\}$ associated with the steady-state covariance matrices $Z = CPC^T$ of the performance outputs $z^t = x^t - x^\star$ (top row) and $z^t = Q^{1/2}(x^t - x^\star)$ (bottom row) for algorithms (2) with the parameters provided in Table II for the matrix Q given in (18) with $m \ll L = O(1)$. The horizontal and vertical axes show the eigenvectors $[1\ 0]^T$ and $[0\ 1]^T$ associated with the eigenvalues $\hat{J}(L)$ and $\hat{J}(m)$ (top row) and $\hat{J}'(L)$ and $\hat{J}'(m)$ (bottom row) of the respective output covariance matrices Z.

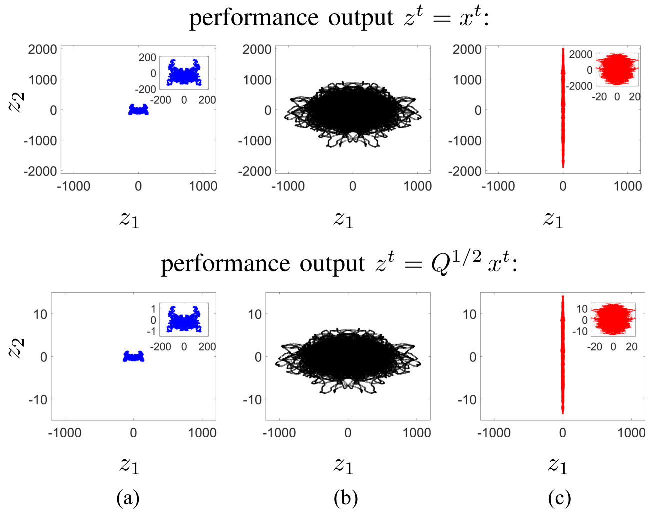

Fig. 2. Performance outputs $z^t = x^t$ (top row) and $z^t = Q^{1/2}x^t$ (bottom row) resulting from 10^5 iterations of noisy first-order algorithms (2) with the parameters provided in Table II. Strongly convex problem with $f(x) = 0.5\,x_1^2 + 0.25 \times 10^{-4}\,x_2^2$ ($\kappa = 2 \times 10^4$) is solved using algorithms with additive white noise and zero initial conditions. (a) Gradient descent. (b) Heavy-ball. (c) Nesterov.

Fig. 2 shows the performance outputs $z^t = x^t$ and $z^t = Q^{1/2}x^t$ resulting from 10^5 iterations of noisy first-order algorithms with the optimal parameters provided in Table II for the strongly convex objective function $f(x) = 0.5\,x_1^2 + 0.25 \times 10^{-4}\,x_2^2$ ($\kappa = 2 \times 10^4$). Although Nesterov's method exhibits good performance with respect to the error in the objective function (performance measure J'), the plots in the first row illustrate detrimental impact of noise on both accelerated algorithms with respect to the variance of the iterates (performance measure J). In particular, we observe that: 1) for gradient descent and the

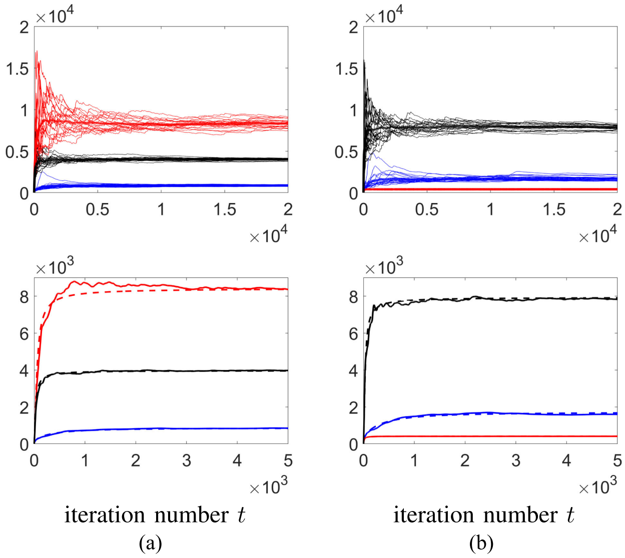

Fig. 3. $(1/t)\sum_{k=0}^{t}\|z^k\|^2$ for the performance output z^t in Example 2. Top row: the thick blue (gradient descent), black (heavy-ball), and red (Nesterov's method) lines mark variance obtained by averaging results of 20 stochastic simulations. Bottom row: comparison between results obtained by averaging outcomes of twenty stochastic simulations (thick lines) with the corresponding theoretical values $(1/t)\sum_{k=0}^{t}\text{trace}\,(CP^kC^T)$ (dashed lines) resulting from the Lyapunov equation (6a). (a) $z^t = x^t$. (b) $z^t = Q^{1/2}x^t$.

heavy-ball method, the iterates x^t are scattered uniformly along the eigendirections of the Hessian matrix Q and acceleration increases variance equally along all directions; and 2) relative to gradient descent, Nesterov's method exhibits larger variance in the iterates x^t along the direction that corresponds to the smallest eigenvalue $\lambda_{\min}(Q)$.

Example 2: Fig. 3 compares the results of 20 stochastic simulations for a strongly convex quadratic objective function (4) with $q = 0$ and a Toeplitz matrix $Q \in \mathbb{R}^{50\times 50}$ with the first row $[2\ -1\ 0\cdots 0\ 0]^T$. This figure shows the time dependence of the variance of the performance outputs $z^t = x^t$ and $z^t = Q^{1/2}x^t$ for the algorithms subject to additive white noise with zero initial conditions. The plots further demonstrate that the mean error in the objective function does not capture detrimental impact of noise on the variance of the iterates for Nesterov's algorithm. The bottom row also compares variance obtained by averaging outcomes of twenty stochastic simulations with the corresponding theoretical values resulting from the Lyapunov equations.

IV. General Strongly Convex Problems

In this section, we extend our results to the class $\mathcal{F}_m^L$ of m-strongly convex objective functions with L-Lipschitz continuous gradients. While a precise characterization of noise amplification for general problems is challenging because of the nonlinear dynamics, we employ tools from robust control theory to obtain meaningful upper bounds. Our results utilize the theory of integral quadratic constraints [51], a convex control-theoretic framework that was recently used to analyze optimization algorithms [50] and study convergence and robustness of the first-order methods [52]–[55]. We establish analytical upper bounds on the mean-squared error of the iterates (3) for gradient descent (2a) and Nesterov's accelerated (2c) methods. Since there are no known accelerated convergence guarantees for the heavy-ball method when applied to general strongly convex functions, we do not consider it in this section.

We first exploit structural properties of the gradient and employ quadratic Lyapunov functions to formulate a semidefinite programming problem (SDP) that provides upper bounds on J in (3). While quadratic Lyapunov functions yield tight upper bounds for gradient descent, they fail to provide any upper bound for Nesterov's method for large condition numbers ($\kappa > 100$). To overcome this challenge, we present a modified semidefinite program that uses more general Lyapunov functions, which are obtained by augmenting standard quadratic terms with the objective function. This type of generalized Lyapunov functions has been introduced in [53] and [56] and used to study convergence of optimization algorithms for nonstrongly convex problems. We employ a modified SDP to derive meaningful upper bounds on J in (3) for Nesterov's method as well.

We note that algorithms (2) are invariant under translation, i.e., if we let $\tilde{x} := x - \bar{x}$ and $g(\tilde{x}) := f(\tilde{x} + \bar{x})$, then (2c), for example, satisfies

$$\begin{aligned}\tilde{x}^{t+2} = {} & \tilde{x}^{t+1} + \beta(\tilde{x}^{t+1} - \tilde{x}^t) - \\ & \alpha\nabla g\left(\tilde{x}^{t+1} + \beta(\tilde{x}^{t+1} - \tilde{x}^t)\right) + \sigma w^t.\end{aligned}$$

Thus, in what follows, without loss of generality, we assume that $x^\star = 0$ is the unique minimizer of (1).

A. Approach Based on Contraction Mappings

Before we present our approach based on LMIs, we provide a more intuitive approach that can be used to examine noise amplification of gradient descent. Let $\varphi\colon \mathbb{R}^n \to \mathbb{R}^n$ be a contraction mapping, i.e., there exists a positive scalar $\eta < 1$ such that $\|\varphi(x) - \varphi(y)\| \le \eta\|x - y\|$ for all $x, y \in \mathbb{R}^n$, and let $x^\star = 0$ be the unique fixed point of φ, i.e, $\varphi(0) = 0$. For the noisy recursion $x^{t+1} = \varphi(x^t) + \sigma w^t$, where w^t is a zero-mean white noise with identity covariance and $\mathbb{E}((w^t)^T\varphi(x^t)) = 0$, the contractiveness of φ implies

$$\mathbb{E}(\|x^{t+1}\|^2) = \mathbb{E}(\|\varphi(x^t) + \sigma w^t\|^2) \le \eta^2\,\mathbb{E}(\|x^t\|^2) + n\sigma^2.$$

Since $\eta < 1$, this relation yields

$$\lim_{t\to\infty}\mathbb{E}(\|x^t\|^2) \le \frac{n\sigma^2}{1-\eta^2}.$$

If $\eta := \max\{|1-\alpha m|, |1-\alpha L|\} < 1$, the map $\varphi(x) := x - \alpha\nabla f(x)$ is a contraction [57]. Thus, for the conventional stepsize $\alpha = 1/L$ we have $\eta = 1 - 1/\kappa$, and the bound becomes

$$\lim_{t\to\infty}\mathbb{E}(\|x^t\|^2) \le \frac{n\sigma^2}{1-\eta^2} = \frac{n\sigma^2\kappa^2}{2\kappa - 1} = n\Theta(\kappa).$$

In the next section, we show that this upper bound is indeed tight for the class of functions $\mathcal{F}_m^L$. While this approach yields a tight upper bound for gradient descent, it cannot be used for Nesterov's method (because it is not a contraction).

B. Approach Based on LMIs

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

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

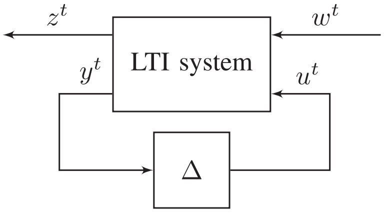

Fig. 4. Block diagram of system (22a).

satisfies the quadratic inequality [50, Lemma 6]

$$\begin{bmatrix} y \,-\, y_0 \\ \Delta(y) \,-\, \Delta(y_0) \end{bmatrix}^T \Pi \begin{bmatrix} y \,-\, y_0 \\ \Delta(y) \,-\, \Delta(y_0) \end{bmatrix} \geq 0 \tag{20}$$

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

$$\Pi \;:=\; \begin{bmatrix} 0 & (L \,-\, m)I \\ (L \,-\, m)I & -2I \end{bmatrix}. \tag{21}$$

We can bring algorithms (2) with constant parameters into a time-invariant state-space form

$$\begin{aligned} \psi^{t+1} &= A\psi^t + \sigma B_w w^t + B_u u^t \\ \begin{bmatrix} z^t \\ y^t \end{bmatrix} &= \begin{bmatrix} C_z \\ C_y \end{bmatrix} \psi^t \\ u^t &= \Delta(y^t) \end{aligned} \tag{22a}$$

that contains a feedback interconnection of linear and nonlinear components. Fig. 4 illustrates the block diagram of system (22a), where ψ^t is the state, w^t is a white stochastic noise, z^t is the performance output, and u^t is the output of the nonlinear term $\Delta(y^t)$. In particular, if we let

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

and define the corresponding matrices as

$$\begin{aligned} A &= \begin{bmatrix} 0 & I \\ -\beta(1-\alpha m)I & (1+\beta)(1-\alpha m)I \end{bmatrix} \\ B_w &= \begin{bmatrix} 0 \\ I \end{bmatrix}, B_u = \begin{bmatrix} 0 \\ -\alpha I \end{bmatrix} \\ C_z &= [\,I \quad 0\,], \quad C_y = [\,-\beta\, I \quad (1+\beta)I\,] \end{aligned} \tag{22b}$$

then (22a) represents Nesterov's method (2c). For gradient descent (2a), we can alternatively use $\psi^t = z^t = y^t := x^t$ with the corresponding matrices

$$\begin{aligned} A &= (1-\alpha\, m)I, \quad B_w = I, \quad B_u = -\alpha I \\ C_z &= C_y = I. \end{aligned} \tag{22c}$$

In what follows, we demonstrate how property (20) of the nonlinear mapping Δ allows us to obtain upper bounds on J when system (22a) is driven by the white stochastic input w^t with zero mean and identity covariance. Lemma 1 uses a quadratic Lyapunov function of the form $V(\psi) = \psi^T X \psi$ and provides upper bounds on the steady-state second-order moment of the performance output z^t in terms of solutions to a certain LMI. This approach yields a tight upper bound for gradient descent.

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

$$\begin{bmatrix} y \\ u \end{bmatrix}^T \Pi \begin{bmatrix} y \\ u \end{bmatrix} \geq 0 \tag{23}$$

for some matrix Π, let X be a positive-semidefinite matrix, and let λ be a nonnegative scalar such that system (22a) satisfies

$$\begin{aligned} &\begin{bmatrix} A^T X\, A - X + C_z^T\, C_z & A^T X\, B_u \\ B_u^T\, X\, A & B_u^T\, X\, B_u \end{bmatrix} + \\ &\lambda \begin{bmatrix} C_y^T & 0 \\ 0 & I \end{bmatrix} \Pi \begin{bmatrix} C_y & 0 \\ 0 & I \end{bmatrix} \preceq \quad 0. \end{aligned} \tag{24}$$

Then, the steady-state second-order moment J of the performance output z^t in (22a) is bounded by

$$J \leq \sigma^2 \text{trace}\,(B_w^T X B_w).$$

Proof: See Appendix B. ∎

For Nesterov's accelerated method with the parameters provided in Table I, computational experiments show that LMI (24) becomes infeasible for large values of the condition number κ. Thus, Lemma 1 does not provide sensible upper bounds on J for Nesterov's algorithm. This observation is consistent with the results of [50], where it was suggested that analyzing the convergence rate requires the use of additional quadratic inequalities, apart from (20), to further tighten the constraints on the gradient ∇f and reduce conservativeness. In what follows, we build on the results of [53] and present an alternative LMI in Lemma 2 that is obtained using a Lyapunov function of the form $V(\psi) = \psi^T X \psi + f([\,0 I\,]\psi)$, where X is a positive-semidefinite matrix and f is the objective function in (1). Such Lyapunov functions have been used to study convergence of optimization algorithms in [56]. The resulting approach allows us to establish an orderwise tight analytical upper bound on J for Nesterov's accelerated method.

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 (22a), (22b) for algorithm (2c), and let Π be given by (21). Then, for any positive-semidefinite matrix X and scalars $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$ that satisfy

$$\begin{aligned} &\begin{bmatrix} A^T X\, A - X + C_z^T\, C_z & A^T X\, B_u \\ B_u^T\, X\, A & B_u^T\, X\, B_u \end{bmatrix} + \\ &\lambda_1 \begin{bmatrix} C_y^T & 0 \\ 0 & I \end{bmatrix} \Pi \begin{bmatrix} C_y & 0 \\ 0 & I \end{bmatrix} \,+\, \lambda_2\, M \qquad \preceq 0 \end{aligned} \tag{25}$$

the steady-state second-order moment J of the performance output z^t in (22a) is bounded by

$$J \leq \sigma^2 \left(n\, L\, \lambda_2 + \text{trace}\,(B_w^T X B_w) \right). \tag{26}$$

Proof: See Appendix B. ∎

Remark 3: Since LMI (25) simplifies to (24) by setting $\lambda_2 = 0$, Lemma 2 represents a relaxed version of Lemma 1. This modification is the key enabler to establishing tight upper bound on J for Nesterov's method.

The upper bounds provided in Lemmas 1 and 2 are proportional to σ^2. In what follows, to make a connection between these bounds and our analytical expressions for the variance amplification in the quadratic case (see Section III), we again set $\sigma = 1$. The best upper bound on J that can be obtained using Lemma 2 is given by the optimal objective value of the semidefinite program

$$\begin{aligned} &\underset{X,\lambda_1,\lambda_2}{\text{minimize}} \quad n\, L\, \lambda_2 \;+\; \text{trace}\,(B_w^T X B_w) \\ &\text{subject to} \quad \text{LMI}\,(25), X \succeq 0, \lambda_1 \geq 0, \lambda_2 \geq 0. \end{aligned} \tag{27}$$

For system matrices (22b), LMI (25) is of size $3n \times 3n$, where $x^t \in \mathbb{R}^n$. However, if we impose the additional constraint that the matrix X has the same block structure as A

$$X = \begin{bmatrix} x_1 I & x_0 I \\ x_0 I & x_2 I \end{bmatrix}$$

for some scalars x_1, x_2, and x_0, then using appropriate permutation matrices, we can simplify (24) into an LMI of size 3×3. Furthermore, imposing this constraint comes without loss of generality. In particular, the optimal objective value of problem (27) does not change if we require X to have this structure; see [50, Sec. 4.2] for a discussion of this lossless dimensionality reduction for LMI constraints with similar structure.

In Theorem 5, we use Lemmas 1 and 2 to establish tight upper bounds on J_{gd} and J_{na} for all $f \in \mathcal{F}_m^L$.

Theorem 5: For gradient descent and Nesterov's accelerated method with the parameters provided in Table I and $\sigma = 1$, the performance measures J_{gd} and J_{na} of the error $x^t - x^\star \in \mathbb{R}^n$ satisfy

$$\sup_{f \in \mathcal{F}_m^L} J_{\text{gd}} \;=\; q_{\text{gd}}, \qquad q_{\text{na}} \;\leq\; \sup_{f \in \mathcal{F}_m^L} J_{\text{na}} \;\leq 4.08\, q_{\text{na}}$$

where

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

and $\kappa := L/m$ is the condition number of the set $\mathcal{F}_m^L$.

Proof: See Appendix B. ■

The variance amplification of gradient descent and Nesterov's method for $f(x) = \frac{m}{2} x^T x$ in $\mathcal{F}_m^L$ is determined by q_{gd} and q_{na}, respectively, and these two quantities can be obtained using Theorem 1. In Theorem 5, we use this strongly convex quadratic objective function to certify the accuracy of the upper bounds on $\sup J$ for all $f \in \mathcal{F}_m^L$. In particular, we observe that the upper bound is exact for gradient descent and that it is within a 4.08 factor of the optimal for Nesterov's method.

For strongly convex objective functions with the condition number κ, Theorem 5 proves that 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.

V. Tuning of Algorithmic Parameters

The parameters provided in Table II yield the optimal convergence rate for strongly convex quadratic problems. For these specific values, Theorem 4 establishes upper and lower bounds on the variance amplification that reveal the negative impact of acceleration. However, it is relevant to examine whether the parameters can be designed to provide acceleration while reducing the variance amplification.

While the convergence rate solely depends on the extreme eigenvalues $m = \lambda_{\min}(Q)$ and $L = \lambda_{\max}(Q)$ of the Hessian matrix Q, variance amplification is influenced by the entire spectrum of Q and its minimization is challenging as it requires the use of all eigenvalues. In this section, we first consider the special case of eigenvalues being symmetrically distributed over the interval $[m, L]$ and demonstrate that for gradient descent and the heavy-ball method, the parameters provided in Table II yield a variance amplification that is within a constant factor of the optimal value. As we demonstrate in Section VI, symmetric distribution of the eigenvalues is encountered in distributed consensus over undirected torus networks. We also consider the problem of designing parameters for objective functions in which the problem size satisfies $n \ll \kappa$ and establish a tradeoff between convergence rate and variance amplification. More specifically, we show that for a bounded problem dimension n and any accelerating pair of parameters α and β, i.e., α and β for which the corresponding rate of convergence satisfies $\rho = 1 - c/\sqrt{\kappa}$ for some constant c, the variance amplification of accelerated methods is larger than that of gradient descent by a factor of $\Omega(\sqrt{\kappa})$.

A. Tuning of Parameters Using the Whole Spectrum

Let $L = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n = m > 0$ be the eigenvalues of the Hessian matrix Q of the strongly convex quadratic objective function in (4). Algorithms (2) converge linearly in the expected value to the optimizer $x^\star$ with the rate

$$\rho := \max_i \hat{\rho}(\lambda_i) \tag{28}$$

where $\hat{\rho}(\lambda_i)$ is the spectral radius of the matrix $\hat{A}_i$ given by (8). For any scalar $c > 0$ and fixed σ, let

$$\begin{aligned} (\alpha_{\text{hb}}^\star(c), \beta_{\text{hb}}^\star(c)) := \underset{\alpha,\,\beta}{\text{argmin}} \quad & J_{\text{hb}}(\alpha, \beta) \\ \text{subject to} \quad & \rho_{\text{hb}} \leq 1 - \frac{c}{\sqrt{\kappa}} \end{aligned} \tag{29a}$$

for the heavy-ball method, and

$$\begin{aligned} \alpha_{\text{gd}}^\star(c) := \underset{\alpha}{\text{argmin}} \quad & J_{\text{gd}}(\alpha) \\ \text{subject to} \quad & \rho_{\text{gd}} \leq 1 - \frac{c}{\kappa} \end{aligned} \tag{29b}$$

for gradient descent, where the expression for the variance amplification J is provided in Theorem 1. Here, the constraints enforce a standard rate of linear convergence for gradient descent and an accelerated rate of linear convergence for the heavy-ball method parameterized with the constant c. Obtaining a closed-form solution to (29) is challenging because J depends on all eigenvalues of the Hessian matrix Q. Herein, we focus on objective functions for which the spectrum of Q is symmetric, i.e., for any eigenvalue λ, the corresponding mirror image $\lambda' := L + m - \lambda$ with respect to $\frac{1}{2}(L + m)$ is also an eigenvalue with the same algebraic multiplicity. For this class of problems, Theorem 6 demonstrates that the parameters provided in Table II for gradient descent and the heavy-ball method yield variance amplification that is within a constant factor of the optimal.

Theorem 6: For any scalar $c > 0$ and fixed σ, there exist constants $c_1 \geq 1$ and $c_2 > 0$ such that for any strongly convex quadratic objective function in which the spectrum of the Hessian matrix Q is symmetrically distributed over the interval $[m, L]$ with $\kappa := L/m > c_1$, we have

$$\begin{aligned} J_{\mathrm{gd}}(\alpha_{\mathrm{gd}}^{\star}(c)) \;&\geq \frac{1}{2}\, J_{\mathrm{gd}}(\alpha_{\mathrm{gd}}) \\ J_{\mathrm{hb}}(\alpha_{\mathrm{hb}}^{\star}(c), \beta_{\mathrm{hb}}^{\star}(c)) \;&\geq c_2\, J_{\mathrm{hb}}(\alpha_{\mathrm{hb}}, \beta_{\mathrm{hb}}) \end{aligned}$$

where parameters α_{gd} and $(\alpha_{\mathrm{hb}}, \beta_{\mathrm{hb}})$ are provided in Table II, whereas $\alpha_{\mathrm{gd}}^{\star}(c)$ and $(\alpha_{\mathrm{hb}}^{\star}(c), \beta_{\mathrm{hb}}^{\star}(c))$ solve (29).

Proof: See Appendix C. ■

For strongly convex quadratic objective functions with symmetric spectrum of the Hessian matrix over the interval $[m, L]$, Theorem 6 shows that the variance amplifications of gradient descent and the heavy-ball method with the parameters provided in Table II are within constant factors of the optimal values. As we illustrate in Section VI, this class of problems is encountered in distributed averaging over noisy undirected networks. Combining this result with the lower bound on $J_{\mathrm{hb}}(\alpha_{\mathrm{hb}}, \beta_{\mathrm{hb}})$ and the upper bound on $J_{\mathrm{gd}}(\alpha_{\mathrm{gd}})$ established in Theorem 4, we see that regardless of the choice of parameters, there is a fundamental gap of $\Omega(\sqrt{\kappa})$ between J_{hb} and J_{gd} as long as we require an accelerated rate of convergence.

B. Fundamental Lower Bounds

We next establish lower bounds on the variance amplification of accelerated methods that hold for any pair of α and β for strongly convex quadratic problems with $\kappa \gg 1$. In particular, we show that the variance amplification of accelerated algorithms is lower bounded by $\Omega(\kappa^{3/2})$ irrespective of the choice of α and β.

The next theorem establishes a fundamental tradeoff between the convergence rate and variance amplification for the heavy-ball method.

Theorem 7: For strongly convex quadratic problems with any stabilizing parameters $\alpha > 0$ and $0 < \beta < 1$ and with a fixed noise magnitude σ, the heavy-ball method with the linear convergence rate ρ satisfies

$$\frac{J_{\mathrm{hb}}}{1-\rho} \geq \sigma^2 \left(\frac{\kappa+1}{8}\right)^2.$$

Furthermore, if $\sigma = \alpha$, i.e., when the only source of uncertainty is a noisy gradient, we have

$$\frac{J_{\mathrm{hb}}}{1-\rho} \geq \left(\frac{\kappa}{8L}\right)^2.$$

Proof: See [58]. ■

To gain additional insight, let us consider two special cases: 1) for $\alpha = 1/L$ and $\beta \to 0^+$, we obtain a gradient descent algorithm, for which $1-\rho = \Theta(1/\kappa)$ and $J = \Theta(\kappa)$; and 2) for the heavy-ball method with the parameters provided in Table II, we have $1-\rho = \Theta(1/\sqrt{\kappa})$ and $J = \Theta(\kappa\sqrt{\kappa})$. Thus, in both cases, $J_{\mathrm{hb}}/(1-\rho) = \Omega(\kappa^2)$. Theorem 7 shows that this lower bound is fundamental, and it therefore quantifies the tradeoff between the convergence rate and the variance amplification of the heavy-ball method for any choice of parameters α and β. It is also worth noting that the lower bound for $\sigma = \alpha$ depends on the largest eigenvalue L of the Hessian matrix Q. Thus, this bound is meaningful when the value of L is uniformly upper bounded. This scenario occurs in many applications, including consensus over undirected tori networks (see Section VI).

While we are not able to show a similar lower bound for Nesterov's method, in the next theorem, we establish an asymptotic lower bound on the variance amplification that holds for any pair of accelerating parameters (α, β) for both Nesterov's and heavy-ball methods.

Theorem 8: For a strongly convex quadratic objective function with condition number κ, let $c > 0$ be a constant such that either Nesterov's algorithm or the heavy-ball method with some (possibly problem dependent) parameters $\alpha > 0$ and $0 < \beta < 1$ converges linearly with a rate $\rho \leq 1 - c/\sqrt{\kappa}$. Then, for any fixed noise magnitude σ, we have

$$J/\sigma^2 \;=\; \Omega(\kappa^{\frac{3}{2}}).$$

Furthermore, if $\sigma = \alpha$, i.e., when the only source of uncertainty is a noisy gradient, we have

$$J \;=\; \Omega\left(\kappa^{\frac{3}{2}}/L^2\right).$$

Proof: For the heavy-ball method, the result follows from combining Theorem 7 with the inequality $1-\rho \geq c/\sqrt{\kappa}$. For Nesterov's method, see [58]. ■

For problems with $n \ll \kappa$, we recall that the variance amplification of gradient descent with conventional values of parameters scales as $O(\kappa)$ (see Theorem 5). Irrespective of the choice of α and β, this result in conjunction with Theorem 8 demonstrates that acceleration cannot be achieved without increasing the variance amplification J by a factor of $\Omega(\sqrt{\kappa})$.

VI. Application to Distributed Computation Over Undirected Networks

Distributed computation over networks has received significant attention in optimization, control systems, signal processing, communications, and machine learning communities. In this problem, the goal is to optimize an objective function (e.g., for the purpose of training a model) using multiple processing units that are connected over a network. Clearly, the structure of the network (e.g., node dynamics and network topology) may impact the performance (e.g., convergence rate and noise amplification) of any optimization algorithm. As a first step toward understanding the impact of the network structure on performance of noisy first-order optimization algorithms, in this section, we examine the standard distributed consensus problem.

The consensus problem arises in applications ranging from social networks, to distributed computing networks, to cooperative control in multiagent systems. In the simplest setup, each node updates a scalar value using the values of its neighbors such that they all agree on a single consensus value. Simple updating strategies of this kind can be obtained by applying a first-order algorithm to the convex quadratic problem

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} x^T \mathbf{L}\, x \tag{30}$$

where $\mathbf{L} = \mathbf{L}^T \in \mathbb{R}^{n \times n}$ is the Laplacian matrix of the graph associated with the underlying undirected network and $x \in \mathbb{R}^n$ is the vector of node values.

The graph Laplacian matrix $\mathbf{L} \succeq 0$ has a nontrivial null space that consists of the minimizers of problem (30). In the absence of noise, for gradient descent and both of its accelerated variants, it is straightforward to verify that the projections v^t of the iterates x^t onto the null space of $\mathbf{L}$ remain constant ($v^t = v^0$, for all t) and also that x^t converges linearly to v^0. In the presence of additive noise, however, v^t experiences a random walk, which leads to an unbounded variance of x^t as $t \to \infty$. Instead, as described in [38], the performance of algorithms in this case can be quantified by examining $\bar{J} := \lim_{t\to\infty} \mathbb{E}(\|x^t - v^t\|^2)$. For connected networks, the null space of $\mathbf{L}$ is given by $\mathcal{N}(\mathbf{L}) = \{c\mathbb{1} \,|\, c \in \mathbb{R}\}$ and

$$\bar{J} = \lim_{t\to\infty} \mathbb{E}\left(\|x^t - (\mathbb{1}^T x^t/n)\mathbb{1}\|^2\right) \tag{31}$$

quantifies the mean-squared deviation from the network average, where $\mathbb{1}$ denotes the vector of all ones, i.e., $\mathbb{1} := [1 \cdots 1]^T$. Finally, it is straightforward to show that $\bar{J}$ can also be computed using the formulas in Theorem 1 by summing over the nonzero eigenvalues of $\mathbf{L}$.

In what follows, we consider a class of networks whose structure allows for the explicit evaluation of the eigenvalues of the Laplacian matrix $\mathbf{L}$. For d-dimensional torus networks, fundamental performance limitations of standard consensus algorithms in continuous time were established in [39], but it remains an open question whether gradient descent and its accelerated variants suffer from these limitations. We utilize such torus networks to demonstrate that standard gradient descent exhibits the same scaling trends as consensus algorithms studied in [39] and that, in lower spatial dimensions, acceleration always increases variance amplification.

A. Explicit Formulas for d-Dimensional Torus Networks

We next examine the asymptotic scaling trends of the performance metric $\bar{J}$ given by (31) for large problem dimensions $n \gg 1$ and highlight the subtle influence of the distribution of the eigenvalues of $\mathbf{L}$ on the variance amplification for d-dimensional torus networks. Tori with nearest neighbor interactions generalize one-dimensional rings to higher spatial dimensions. Let $\mathbb{Z}_{n_0}$ denote the group of integers modulo n_0. A d-dimensional torus $\mathbb{T}^d_{n_0}$ consists of $n := n_0^d$ nodes denoted by v_a where $a \in \mathbb{Z}^d_{n_0}$ and the set of edges $\{\{v_a v_b\} | \|a - b\| = 1 \mod n_0\}$; nodes v_a and v_b are neighbors if and only if a and b differ exactly at a single entry by one. For example, $\mathbb{T}^1_{n_0}$ denotes a ring with $n = n_0$ nodes and $\mathbb{T}^5_{n_0}$ denotes a five-dimensional torus with $n = n_0^5$ nodes.

The multidimensional discrete Fourier transform can be used to determine the eigenvalues of the Laplacian matrix $\mathbf{L}$ of a d-dimensional torus $\mathbb{T}^d_{n_0}$ as

$$\lambda_i = \sum_{l=1}^{d} 2\left(1 - \cos\tfrac{2\pi i_l}{n_0}\right), i_l \in \mathbb{Z}_{n_0} \tag{32}$$

where $i := (i_1, \ldots, i_d) \in \mathbb{Z}^d_{n_0}$. We note that $\lambda_0 = 0$ is the only zero eigenvalue of $\mathbf{L}$ with the eigenvector $\mathbb{1}$ and that all other eigenvalues are positive. Let $\kappa := \lambda_{\max}/\lambda_{\min}$ be the ratio of the largest and smallest nonzero eigenvalues of $\mathbf{L}$. A key observation is that, for $n_0 \gg 1$,

$$\kappa = \Theta\left(\frac{2}{1 - \cos(2\pi/n_0)}\right) = \Theta(n_0^2) = \Theta(n^{2/d}). \tag{33}$$

This is because $\lambda_{\min} = 2d(1 - \cos(2\pi/n_0))$ goes to zero as $n_0 \to \infty$, and the largest eigenvalue of $\mathbf{L}$, $\lambda_{\max} = 2d(1 - \cos(2\pi\lfloor\frac{n_0}{2}\rfloor/n_0))$, is equal to $4d$ for even n_0, and it approaches $4d$ from below for odd n_0.

As aforementioned, the performance metric $\bar{J}$ can be obtained by

$$\bar{J} = \sum_{0 \neq i \in \mathbb{Z}^d_{n_0}} \hat{J}(\lambda_i)$$

where $\hat{J}(\lambda)$ for each algorithm is determined in Theorem 1 and λ_i are the nonzero eigenvalues of $\mathbf{L}$. The next theorem characterizes the asymptotic value of the network-size normalized mean-squared deviation from the network average, $\bar{J}/n$, for a fixed spatial dimension d and condition number $\kappa \gg 1$. This result is obtained using analytical expression (32) for the eigenvalues of the Laplacian matrix $\mathbf{L}$.

Theorem 9: Let $\mathbf{L} \in \mathbb{R}^{n\times n}$ be the graph Laplacian of the d-dimensional undirected torus $\mathbb{T}^d_{n_0}$ with $n = n_0^d \gg 1$ nodes. For convex quadratic optimization problem (30), the network-size normalized performance metric $\bar{J}/n$ of noisy first-order algorithms, with the parameters provided in Table II and $\sigma = 1$, is determined by

	$d=1$	$d=2$	$d=3$	$d=4$	$d=5$
GD	$\Theta(\sqrt{\kappa})$	$\Theta(\log\,\kappa)$	$\Theta(1)$	$\Theta(1)$	$\Theta(1)$
NA	$\Theta(\kappa)$	$\Theta(\sqrt{\kappa}\log\,\kappa)$	$\Theta(\kappa^{\frac{1}{4}})$	$\Theta(\log\,\kappa)$	$\Theta(1)$
HB	$\Theta(\kappa)$	$\Theta(\sqrt{\kappa}\,\log\,\kappa)$	$\Theta(\sqrt{\kappa})$	$\Theta(\sqrt{\kappa})$	$\Theta(\sqrt{\kappa})$

where $\kappa = \Theta(n^{2/d})$ is the condition number of $\mathbf{L}$ given in (33).

Proof: See Appendix D. ■

Theorem 9 demonstrates that the variance amplification of gradient descent is equivalent to that of the standard consensus algorithm studied in [39] and that, in lower spatial dimensions, acceleration always negatively impacts the performance of noisy algorithms. Our results also highlight the subtle influence of the distribution of the eigenvalues of $\mathbf{L}$ on the variance amplification. For rings (i.e., $d = 1$), lower bounds provided in Theorem 4 capture the trends that our detailed analysis based on the distribution of the entire spectrum of $\mathbf{L}$ reveals. In higher spatial dimensions, however, the lower bounds that are obtained using only the extreme eigenvalues of $\mathbf{L}$ are conservative. Similar conclusion can be made about the upper bounds provided in Theorem 4. This observation demonstrates that the naïve bounds that result only from the use of the extreme eigenvalues can be overly conservative.

We also note that gradient descent significantly outperforms Nesterov's accelerated algorithm in lower spatial dimensions. In particular, while $\bar{J}/n$ becomes network size independent for $d = 3$ for gradient descent, Nesterov's algorithm reaches "critical connectivity" only for $d = 5$. In contrast, in any spatial dimension, there is no network-size-independent upper bound on $\bar{J}/n$ for the heavy-ball method. These conclusions could not have been reached without performing an in-depth analysis of the impact of all eigenvalues on performance of noisy networks with $n \gg 1$ and $\kappa \gg 1$.

VII. Concluding Remarks

We study the robustness of noisy first-order algorithms for smooth, unconstrained, strongly convex optimization problems.

Even though the underlying dynamics of these algorithms are in general nonlinear, we establish upper bounds on noise amplification that are accurate up to constant factors. For quadratic objective functions, we provide analytical expressions that quantify the effect of all eigenvalues of the Hessian matrix on variance amplification. We use these expressions to establish lower bounds demonstrating that although the acceleration techniques improve the convergence rate, they significantly amplify noise for problems with large condition numbers. In problems of bounded dimension $n \ll \kappa$, the noise amplification increases from $O(\kappa)$ to $\Omega(\kappa^{3/2})$ when moving from standard gradient descent to accelerated algorithms. We specialize our results to the problem of distributed averaging over noisy undirected networks and also study the role of network size and topology on robustness of accelerated algorithms. Future research directions include 1) extension of our analysis to multiplicative and correlated noise and 2) robustness analysis of broader classes of optimization algorithms.

Appendix A
Quadratic Problems

Proof of Theorem 1: For gradient descent, $\hat{A}_i = 1 - \alpha\lambda_i$ and $\hat{B}_i = 1$ are scalars, and the solution to (9) is given by

$$\hat{P}_i := \sigma^2 p_i = \frac{\sigma^2}{1 - (1-\alpha\lambda_i)^2} = \frac{\sigma^2}{\alpha\lambda_i(2-\alpha\lambda_i)}.$$

For the accelerated methods, we note that for any $\hat{A}_i$ and $\hat{B}_i$ of the form

$$\hat{A}_i = \begin{bmatrix} 0 & 1 \\ a_i & b_i \end{bmatrix}, \hat{B}_i = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

the solution $\hat{P}_i$ to Lyapunov equation (9) is given by

$$\hat{P}_i = \sigma^2 \begin{bmatrix} p_i & b_i p_i/(1-a_i) \\ b_i p_i/(1-a_i) & p_i \end{bmatrix}$$

where

$$p_i := \frac{a_i - 1}{(a_i+1)(b_i+a_i-1)(b_i-a_i+1)}. \tag{34}$$

The parameters a_i and b_i for Nesterov's algorithm are $\{a_i = -\beta(1-\alpha\lambda_i);\ b_i = (1+\beta)(1-\alpha\lambda_i)\}$, and for the heavy-ball method, we have $\{a_i = -\beta;\ b_i = 1+\beta-\alpha\lambda_i\}$. Now, since $\hat{C}_i = 1$ for gradient descent and $\hat{C}_i = [1\ 0]$ for the accelerated algorithms, it follows that for all three algorithms, we have $\hat{J}(\lambda_i) := \text{trace}\,(\hat{C}_i\hat{P}_i\hat{C}_i^T) = \sigma^2 p_i$. Finally, if we use the expression for p_i for gradient descent and substitute for a_i and b_i in (34) for the accelerated algorithms, we obtain the expressions for $\hat{J}$ in the statement of the theorem. ■

Proof of Proposition 1: See [58]. ■

Proof of Theorem 3: From Proposition 1, it follows that

$$\frac{\hat{J}_{\text{na}}(L)}{\hat{J}_{\text{gd}}(L)} \le \frac{\hat{J}_{\text{na}}(\lambda_i)}{\hat{J}_{\text{gd}}(\lambda_i)} \le \frac{\hat{J}_{\text{na}}(m)}{\hat{J}_{\text{gd}}(m)} \tag{35a}$$

for all λ_i and

$$\sum_{i=1}^{n-1} \hat{J}_{\text{gd}}(\lambda_i) \le (n-1)\hat{J}_{\text{gd}}(m) = (n-1)\hat{J}_{\text{gd}}(L). \tag{35b}$$

For the upper bound, we have

$$\begin{aligned}\frac{J_{\text{na}}}{J_{\text{gd}}} = \frac{\sum_{i=1}^{n}\hat{J}_{\text{na}}(\lambda_i)}{\sum_{i=1}^{n}\hat{J}_{\text{gd}}(\lambda_i)} &\le \frac{\hat{J}_{\text{na}}(L) + \frac{\hat{J}_{\text{na}}(m)}{\hat{J}_{\text{gd}}(m)}\sum_{i=1}^{n-1}\hat{J}_{\text{gd}}(\lambda_i)}{\hat{J}_{\text{gd}}(L) + \sum_{i=1}^{n-1}\hat{J}_{\text{gd}}(\lambda_i)} \\ &\le \frac{\hat{J}_{\text{na}}(L) + (n-1)\hat{J}_{\text{na}}(m)}{\hat{J}_{\text{gd}}(L) + (n-1)\hat{J}_{\text{gd}}(m)}\end{aligned}$$

where the first inequality follows from (35a). The second inequality can be verified by multiplying both sides with the product of the denominators and using $\hat{J}_{\text{gd}}(m) = \hat{J}_{\text{gd}}(L)$, $\hat{J}_{\text{na}}(m) \ge \hat{J}_{\text{na}}(L)$, and (35b). The lower bound follows from a similar argument. ■

Proof of the bounds in (16): See [58]. ■

Appendix B
General Strongly Convex Problems

Proof of Lemma 1: Let us define the positive-semidefinite function $V(\psi) := \psi^T X \psi$, and let $\eta := [\psi^T u^T]^T$. Using (23) and LMI (24), we can write

$$\begin{aligned}\|z^t\|^2 = {} & (\eta^t)^T \begin{bmatrix} C_z^T C_z & 0 \\ 0 & 0 \end{bmatrix} \eta^t \le -\lambda \begin{bmatrix} y^t \\ u^t \end{bmatrix}^T \Pi \begin{bmatrix} y^t \\ u^t \end{bmatrix} + \\ & (\eta^t)^T \left(\begin{bmatrix} X & 0 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} A^T \\ B_u^T \end{bmatrix} X \begin{bmatrix} A^T \\ B_u^T \end{bmatrix}^T \right) \eta^t \\ \le {} & V(\psi^t) - V(\psi^{t+1}) + 2\sigma(\psi^t)^T A^T X\, B_w\, w^t + \\ & \sigma^2 (w^t)^T B_w^T\, X\, B_w\, w^t + 2\sigma(u^t)^T B_u^T\, X\, B_w\, w^t.\end{aligned}$$

Since w^t is a zero-mean white input with identity covariance, which is independent of u^t and x^t, if we take the average of the above inequality over t and expectation over different realizations of w^t, we obtain

$$\begin{aligned}\frac{1}{\bar{T}}\sum_{t=1}^{\bar{T}} \mathbb{E}\left(\|z^t\|^2\right) \le {} & \frac{1}{\bar{T}}\,\mathbb{E}\left(V(\psi^1) - V(\psi^{\bar{T}+1})\right) + \\ & \sigma^2 \text{trace}\,(B_w^T X B_w).\end{aligned}$$

Therefore, letting $\bar{T} \to \infty$ and using $X \succeq 0$ lead to $J \le \sigma^2\text{trace}\,(B_w^T X B_w)$, which completes the proof. ■

In order to prove Lemma 2, we present a technical lemma, which along the lines of results of [53] provides us with an upper bound on the difference between the objective value at two consecutive iterations.

Lemma 3: Let $f \in \mathcal{F}_m^L$ and $\kappa := L/m$. Then, Nesterov's accelerated method, with the notation introduced in Section IV, satisfies

$$\begin{aligned} f(x^{t+2}) - f(x^{t+1}) \le {} & \frac{1}{2}\left(N_1 \begin{bmatrix} \psi^t \\ u^t \end{bmatrix} + \begin{bmatrix} \sigma w^t \\ 0 \end{bmatrix} \right)^T \\ & \times \begin{bmatrix} L\,I & I \\ I & 0 \end{bmatrix} \left(N_1 \begin{bmatrix} \psi^t \\ u^t \end{bmatrix} + \begin{bmatrix} \sigma w^t \\ 0 \end{bmatrix} \right) \\ & + \frac{1}{2}\left(N_2 \begin{bmatrix} \psi^t \\ u^t \end{bmatrix} \right)^T \begin{bmatrix} -m\,I & I \\ I & 0 \end{bmatrix} \left(N_2 \begin{bmatrix} \psi^t \\ u^t \end{bmatrix} \right) \end{aligned}$$

where N_1 and N_2 are defined in Lemma 2.

Proof: See [58]. ■

Proof of Lemma 2: Let us define the positive-semidefinite function $V(\psi) := \psi^T X \psi$ and let $\eta := [\psi^T\ u^T]^T$. Similar to the first part of the proof of Lemma 1, we can use LMI (25) and inequality (20) to write

$$\begin{aligned}\|z^t\|^2 \le\; & V(\psi^t) \,-\, V(\psi^{t+1}) \,+\, 2\sigma(\psi^t)^T A^T X\, B_w\, w^t \\ & +\, \sigma^2 (w^t)^T B_w^T\, X\, B_w\, w^t \,+\, 2\sigma(u^t)^T B_u^T\, X\, B_w\, w^t \\ & -\, (\eta^t)^T M \eta^t. \end{aligned} \tag{36}$$

From Lemma 3, it follows that

$$\begin{aligned}(\eta^t)^T M \eta^t \;\ge\; & 2\left(f(x^{t+2}) \,-\, f(x^{t+1})\right) \,-\, \sigma^2 L\,\|w^t\|^2 \\ -\, 2 & \begin{bmatrix}\sigma w^t \\ 0\end{bmatrix}^T \begin{bmatrix} L\,I & I \\ I & 0\end{bmatrix} N_1 \eta^t. \end{aligned} \tag{37}$$

Now, combining inequalities (36) and (37) yields

$$\begin{aligned}\|z^t\|^2 \le\; & V(\psi^t) \,-\, V(\psi^{t+1}) \,+\, 2\sigma(\psi^t)^T A^T X\, B_w\, w^t \\ & +\, \sigma^2 (w^t)^T B_w^T\, X\, B_w\, w^t + 2\sigma(u^t)^T B_u^T\, X\, B_w\, w^t \\ & -\, 2\,\lambda_2 \left(f(x^{t+2}) \,-\, f(x^{t+1})\right) \,+\, \lambda_2 \sigma^2 L \|w^t\|^2 \\ & +\, 2\lambda_2 \begin{bmatrix}\sigma w^t \\ 0\end{bmatrix}^T \begin{bmatrix} L\,I & I \\ I & 0\end{bmatrix} N_1 \eta^t. \end{aligned} \tag{38}$$

Since w^t is a zero-mean white input with identity covariance which is independent of u^t and x^t, taking the expectation of the last inequality yields

$$\begin{aligned}\mathbb{E}\left(\|z^t\|^2\right) \le\; & \mathbb{E}\left(V(\psi^t) - V(\psi^{t+1})\right) + \sigma^2 \mathrm{trace}\,(B_w^T\, X\, B_w) \\ & + 2\lambda_2\, \mathbb{E}\left(f(x^{t+1}) - f(x^{t+2})\right) + n\,\sigma^2 L\,\lambda_2 \end{aligned}$$

and taking the average over the first $\bar{T}$ iterations results in

$$\begin{aligned}\frac{1}{\bar{T}} \sum_{t=1}^{\bar{T}} \mathbb{E}\left(\|z^t\|^2\right) \le\; & \frac{1}{\bar{T}}\, \mathbb{E}\left(V(\psi^1) - V(\psi^{\bar{T}+1})\right) \\ + \sigma^2 \mathrm{trace}\,(B_w^T\, X\, B_w) \,+\, & \frac{2\lambda_2}{\bar{T}}\, \mathbb{E}\left(f(x^2) - f(x^{\bar{T}+2})\right) \\ +\; n\,\sigma^2 L\,\lambda_2. & \end{aligned}$$

Finally, using positive-definiteness of the function V, strong convexity of the function f, and letting $\bar{T} \to \infty$, it follows that $J \le \sigma^2(nL\lambda_2 + \mathrm{trace}\,(B_w^T X\, B_w))$, as required. ■

Proof of Theorem 5: Using Theorem (1), it is straightforward to show that for gradient descent and Nesterov's method with the parameters provided in Table I, the function $f(x) := \frac{m}{2}\|x\|^2$ leads to the largest variance amplification J among the quadratic objective functions within $\mathcal{F}_m^L$. This yields the lower bounds

$$q_{\mathrm{gd}} \,=\, J_{\mathrm{gd}} \,\le\, J_{\mathrm{gd}}^\star, \quad q_{\mathrm{na}} \,=\, J_{\mathrm{na}} \,\le\, J_{\mathrm{na}}^\star$$

with J_{gd} and J_{na} corresponding to $f(x) = \frac{m}{2}\|x\|^2$. We next show that $J_{\mathrm{gd}} \le q_{\mathrm{gd}}$.

To obtain the best upper bound on J_{gd} using Lemma 1, we minimize $\mathrm{trace}\,(B_w^T X B_w)$ subject to LMI (24), $X \succeq 0$, and $\lambda \ge 0$. For gradient descent, if we use representation (22c), then the negative-definiteness of the (1,1)-block of LMI (24) implies that

$$X \succeq \frac{1}{\alpha m(2 - \alpha m)} I = \frac{\kappa^2}{2\kappa - 1} I. \tag{39}$$

It is straightforward to show that the pair

$$X \;=\; \frac{\kappa^2}{2\kappa - 1} I, \lambda \;=\; \frac{1 - \alpha\, m}{m(2 - \alpha\, m)(L - m)} \tag{40}$$

is feasible as the LMI (24) becomes

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

Thus, X and λ given by (40) provide a solution to LMI (24). Therefore, inequality (39) is tight, and it provides the best achievable upper bound

$$J_{\mathrm{gd}} \;\le \mathrm{trace}\,(B_w^T X B_w) \;=\; \frac{n\kappa^2}{2\,\kappa - 1}.$$

For Nesterov's method, the proof of $J_{\mathrm{na}} \le 4.08 q_{\mathrm{na}}$ is provided in [58]. ■

Appendix C
Proof of Theorem 6

Without loss of generality, let $\sigma = 1$ and

$$G \;:=\; \sum_{i=1}^{n} \max\{\hat{J}(\lambda_i), \hat{J}(\lambda_i')\} \tag{41}$$

where λ_i are the eigenvalues of the Hessian of the objective function f and $\lambda_i' = m + L - \lambda_i$ is the mirror image of λ_i with respect to $(m + L)/2$. Since $J = \sum_i \hat{J}(\lambda_i)$, if λ_i are symmetrically distributed over the interval $[m, L]$ i.e., $(\lambda_1, \ldots, \lambda_n) = (\lambda_n', \ldots, \lambda_1')$, then for any parameters α and β, we have

$$J \;\le G \;\le 2J. \tag{42}$$

Equation (42) implies that any bound on G simply carries over to J within an accuracy of constant factors. Thus, we focus on G and establish one of its useful properties in the next lemma that allows us to prove Theorem 6.

Lemma 4: The heavy-ball method with any stabilizing parameter β satisfies

$$\frac{2(1 + \beta)}{L + m} \;=\; \underset{\alpha}{\mathrm{argmin}}\, \rho(\alpha, \beta) \tag{43}$$

where ρ is the rate of linear convergence. Furthermore, if the Hessian of the quadratic objective function f has a symmetric spectrum over the interval $[\lambda_1, \lambda_n] = [m, L]$, then

$$\frac{2(1 + \beta)}{L + m} = \underset{\alpha}{\mathrm{argmin}}\; G(\alpha,\, \beta).$$

Proof: See [58]. ■

Since gradient descent is obtained from the heavy-ball method by letting $\beta = 0$, from Lemma 4, it immediately follows that $\alpha_{\mathrm{gd}} = 2/(L + m)$ in Table II optimizes both G_{gd} and the convergence rate ρ_{gd}. This fact combined with (42) yields

$$2\, J_{\mathrm{gd}}(\alpha_{\mathrm{gd}}^\star(c)) \,\ge\, G_{\mathrm{gd}}(\alpha_{\mathrm{gd}}^\star(c)) \,\ge\, G_{\mathrm{gd}}(\alpha_{\mathrm{gd}}) \,\ge\, J_{\mathrm{gd}}(\alpha_{\mathrm{gd}}) \tag{44}$$

where $\alpha_{\mathrm{gd}}^\star(c)$ is given by (29b). This completes the proof for gradient descent. For the heavy-ball method, the proof is provided in [58].

Appendix D
Proof of Theorem 9

The proof uses the explicit expression for the eigenvalues of torus in (32) to compute the variance amplification $\bar{J} = \sum_{i \neq 0} \hat{J}(\lambda_i)$ for all three algorithms. Several technical results that we use in the proof are presented next.

We borrow the following lemma, which provides tight bounds on the sum of reciprocals of the eigenvalues of a d-dimensional torus network, from [39, Appendix B].

Lemma 5: The eigenvalues λ_i of the graph Laplacian of the d-dimensional torus $\mathbb{T}_{n_0}^d$ with $n_0 \gg 1$ satisfy

$$\sum_{0 \neq i \in \mathbb{Z}_{n_0}^d} \frac{1}{\lambda_i} = \Theta(B(n_0))$$

where the function B is given by

$$B(n_0) = \begin{cases} \dfrac{1}{d-2}(n_0^d - n_0^2), & d \neq 2 \\ n_0^d \log n_0, & d = 2. \end{cases}$$

We next use Lemma 5 to establish an asymptotic expression for the variance amplification of the gradient descent algorithm for a d-dimensional torus.

Lemma 6: For the consensus problem over a d-dimensional torus $\mathbb{T}_{n_0}^d$ with $n_0 \gg 1$, the performance metric $\bar{J}_{\text{gd}}$ corresponding to gradient decent with the stepsize $\alpha = 2/(L+m)$ satisfies $\bar{J}_{\text{gd}} = \Theta(B(n_0))$, where the function B is given in Lemma 5.

Proof: Using the expression for the variance amplification of gradient descent from Theorem 1, we have

$$\begin{aligned} \bar{J}_{\text{gd}} &= \sum_{0 \neq i \in \mathbb{Z}_{n_0}^d} \frac{1}{\alpha \lambda_i (2 - \alpha \lambda_i)} = \frac{1}{2\alpha} \sum_{0 \neq i \in \mathbb{Z}_{n_0}^d} \frac{1}{\lambda_i} + \frac{1}{\frac{2}{\alpha} - \lambda_i} \\ &= \frac{1}{2\alpha} \sum_{0 \neq i \in \mathbb{Z}_{n_0}^d} \frac{1}{\lambda_i} + \frac{1}{\lambda_{\max} + \lambda_{\min} - \lambda_i} \\ &\approx \frac{1}{\alpha} \sum_{0 \neq i \in \mathbb{Z}_{n_0}^d} \frac{1}{\lambda_i} \approx 2d \sum_{0 \neq i \in \mathbb{Z}_{n_0}^d} \frac{1}{\lambda_i}. \end{aligned}$$

The first approximation follows from the facts that the eigenvalues satisfy $0 < \lambda_i \leq \lambda_{\max} + \lambda_{\min} \approx 4d$ and that their distribution is asymptotically symmetric with respect to $\lambda = 2d$. The second approximation follows from

$$\alpha = \frac{2}{L+m} = \frac{2}{\lambda_{\max} + \lambda_{\min}} \approx \frac{1}{2d}.$$

The bounds for the sum of reciprocals of λ_i provided in Lemma 5 can now be used to complete the proof. ■

For gradient descent, the proof of Theorem 9 follows from dividing the asymptotic bounds in Lemma 6 with the total number of nodes $n = n_0^d$. For the heavy-ball method, the result follows from the proof for gradient descent, the relationship between variance amplifications of gradient descent and the heavy-ball method in Theorem 2, and (33). For Nesterov's method, the proof is provided in [58].

References

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

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

[3] 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, Jan. 2016.

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

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

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

[7] 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.

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

[9] Y. Nesterov, *Lectures on Convex Optimization*, vol. 137. New York, NY, USA: Springer, 2018.

[10] D. Maclaurin, D. Duvenaud, and R. Adams, "Gradient-based hyperparameter optimization through reversible learning," in *Proc. Int. Conf. Mach. Learn.*, 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. Int. Conf. Neural Inf. Process.*, 2017, pp. 3458–3468.

[13] M. Fazel, R. Ge, S. M. Kakade, and M. Mesbahi, "Global convergence of policy gradient methods for the linear quadratic regulator," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1467–1476.

[14] H. Mohammadi, A. Zare, M. Soltanolkotabi, and M. R. Jovanović, "Convergence and sample complexity of gradient methods for the model-free linear quadratic regulator problem," *IEEE Trans. Autom. Control*, to be published.

[15] H. Mohammadi, M. Soltanolkotabi, and M. R. Jovanović, "Learning the model-free linear quadratic regulator via random search," in *Proc. 2nd Annu. Conf. Learn. Dyn. Control*, 2020, vol. 120, pp. 1–9.

[16] H. Mohammadi, M. Soltanolkotabi, and M. R. Jovanović, "On the linear convergence of random search for discrete-time LQR," *IEEE Control Syst. Lett.*, to be published, doi: 10.1109/LCSYS.2020.3006256.

[17] R. Ge, F. Huang, C. Jin, and Y. Yuan, "Escaping from saddle points—Online stochastic gradient for tensor decomposition," in *Proc. 28th Conf. Learn. Theory*, 2015, vol. 40, pp. 797–842.

[18] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan, "How to escape saddle points efficiently," in *Proc. Int. Conf. Mach. Learn.*, 2017, vol. 70, pp. 1724–1732.

[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.*, vol. 22, 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, Center Oper. Res. Econometrics, Univ. Catholique de Louvain, Louvain-la-Neuve, Belgium, 2013.

[23] O. Devolder, F. Glineur, and Y. Nesterov, "First-order methods of smooth convex optimization with inexact oracle," *Math. Program.*, vol. 146, nos. 1/2, pp. 37–75, 2014.

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

[25] M. Schmidt, N. L. Roux, and F. R. Bach, "Convergence rates of inexact proximal-gradient methods for convex optimization," in *Proc. Int. Conf. Neural Inf. Process.*, 2011, pp. 1458–1466.

[26] O. Devolder, "Stochastic first order methods in smooth convex optimization," Catholic Univ. Louvain, Louvain-la-Neuve, Belgium, CORE Discusssion Paper 2011/70, 2011.

[27] F. 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.

[28] B. T. Polyak, "New stochastic approximation type procedures," *Autom. i Telemekh*, vol. 7, pp. 98–107, 1990.

[29] B. T. Polyak and A. B. Juditsky, "Acceleration of stochastic approximation by averaging," *SIAM J. Control Optim.*, vol. 30, no. 4, pp. 838–855, 1992.

[30] 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.

[31] E. Moulines and F. Bach, "Non-asymptotic analysis of stochastic approximation algorithms for machine learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2011, pp. 451–459.

[32] N. Tripuraneni, N. Flammarion, F. Bach, and M. I. Jordan, "Averaging stochastic gradient descent on Riemannian manifolds," in *Proc. 31st Conf. Learn. Theory*, 2018, pp. 650–687.

[33] M. Baes, "Estimate sequence methods: Extensions and approximations," ETH Zurich, Zurich, Switzerland, IFOR Internal Rep., Aug. 2009.

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

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

[36] B. T. Polyak, *Introduction to Optimization*, vol. 1. New York, NY, USA: Optimization Software Inc., 1987.

[37] H. Kwakernaak and R. Sivan, *Linear Optimal Control Systems*. Hoboken, NJ, USA: Wiley-Interscience, 1972.

[38] L. Xiao, S. Boyd, and S.-J. Kim, "Distributed average consensus with least-mean-square deviation," *J. Parallel Distrib. Comput.*, vol. 67, no. 1, pp. 33–46, 2007.

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

[40] F. Lin, M. Fardad, and M. R. Jovanović, "Optimal control of vehicular formations with nearest neighbor interactions," *IEEE Trans. Autom. Control*, vol. 57, no. 9, pp. 2203–2218, Sep. 2012.

[41] F. Dörfler, M. R. Jovanović, M. Chertkov, and F. Bullo, "Sparsity-promoting optimal wide-area control of power networks," *IEEE Trans. Power Syst.*, vol. 29, no. 5, pp. 2281–2291, Sep. 2014.

[42] J. W. Simpson-Porco, "Input/output analysis of primal-dual gradient algorithms," in *Proc. 54th Annu. Allerton Conf. Commun., Control, Comput.*, 2016, pp. 219–224.

[43] J. W. Simpson-Porco, B. K. Poolla, N. Monshizadeh, and F. Drfler, "Quadratic performance of primal-dual methods with application to secondary frequency control of power systems," in *Proc. 55th IEEE Conf. Decis. Control*, 2016, pp. 1840–1845.

[44] H. Mohammadi, M. Razaviyayn, and M. R. Jovanović, "Variance amplification of accelerated first-order algorithms for strongly convex quadratic optimization problems," in *Proc. 57th IEEE Conf. Decis. Control*, Miami, FL, USA, 2018, pp. 5753–5758.

[45] H. Mohammadi, M. Razaviyayn, and M. R. Jovanović, "Performance of noisy Nesterov's accelerated method for strongly convex optimization problems," in *Proc. Amer. Control Conf.*, Philadelphia, PA, USA, 2019, pp. 3426–3431.

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

[47] N. S. Aybat, A. Fallah, M. Gürbüzbalaban, and A. Ozdaglar, "A universally optimal multistage accelerated stochastic gradient method," in *Proc. Int. Conf. Neural Inf. Process.*, 2019, pp. 8525–8536.

[48] 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.

[49] S. Michalowsky, C. Scherer, and C. Ebenbauer, "Robust and structure exploiting optimization algorithms: An integral quadratic constraint approach," 2019, *arXiv:1905.00279*.

[50] 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.

[51] A. Megretski and A. Rantzer, "System analysis via integral quadratic constraints," *IEEE Trans. Autom. Control*, vol. 42, no. 6, pp. 819–830, Jun. 1997.

[52] B. Hu and L. Lessard, "Dissipativity theory for Nesterov's accelerated method," in *Proc. Int. Conf. Mach. Learn.*, 2017, vol. 70, pp. 1549–1557.

[53] 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.

[54] S. Cyrus, B. Hu, B. V. Scoy, and L. Lessard, "A robust accelerated optimization algorithm for strongly convex functions," in *Proc. Amer. Control Conf.*, 2018, pp. 1376–1381.

[55] N. K. Dhingra, S. Z. Khong, and M. R. Jovanović, "The proximal augmented Lagrangian method for nonsmooth composite optimization," *IEEE Trans. Autom. Control*, vol. 64, no. 7, pp. 2861–2868, Jul. 2019.

[56] B. T. Polyak and P. Shcherbakov, "Lyapunov functions: An optimization theory perspective," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 7456–7461, 2017.

[57] D. P. Bertsekas, *Convex Optimization Algorithms*. Belmont, MA, USA: Athena Scientific, 2015.

[58] H. Mohammadi, M. Razaviyayn, and M. R. Jovanović, "Robustness of accelerated first-order algorithms for strongly convex optimization problems," 2019, *arXiv:1905.11011*.

Hesameddin Mohammadi (Student Member, IEEE) received the B.Sc. degree from the Sharif University of Technology, Tehran, Iran, in 2015, and the M.Sc. degree from Arizona State University, Tempe, AZ, USA, in 2017, both in mechanical engineering. He is currently working toward the Ph.D. degree in electrical engineering with the Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA, USA.

His current research interests include nonconvex optimization, control, and inference problems.

Meisam Razaviyayn received the Ph.D. degree in electrical engineering with minor in computer science, from the University of Minnesota, Minneapolis, MN, USA, in 2014.

He is currently an Assistant Professor with the Department of Industrial and Systems Engineering, University of Southern California (USC), CA, USA. Prior to joining USC, he was a Postdoctoral Research Fellow with the Department of Electrical Engineering, Stanford University, Stanford, CA, USA, working with Prof. D. Tse.

Dr. Razaviyayn is the recipient of the Signal Processing Society Young Author Best Paper Award in 2014 and the finalist for Best Paper Prize for Young Researcher in Continuous Optimization in 2013 and 2016.

Mihailo R. Jovanović (Fellow, IEEE) received the Ph.D. degree in mechanical engineering from the University of California, Santa Barbara, CA, USA, in 2004.

He is currently a Professor with the Ming Hsieh Department of Electrical and Computer Engineering and the Founding Director of the Center for Systems and Control, University of Southern California, CA, USA. He was a Faculty Member with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, USA, from 2004 to 2017. He has held visiting positions with Stanford University and the Institute for Mathematics and its Applications.

Dr. Jovanović is a Fellow of the American Physical Society. He received a CAREER Award from the National Science Foundation in 2007, the George S. Axelby Outstanding Paper Award from the IEEE Control Systems Society in 2013, and the Distinguished Alumni Award from the Department of Mechanical Engineering, University of California, Santa Barbara, in 2014. Papers of his students were finalists for the Best Student Paper Award at the American Control Conference in 2007 and 2014.