2018 IEEE Conference on Decision and Control (CDC)
Miami Beach, FL, USA, Dec. 17-19, 2018

Variance amplification of accelerated first-order algorithms for strongly convex quadratic optimization problems

Hesameddin Mohammadi, Meisam Razaviyayn, and Mihailo R. Jovanović

Abstract— **We study performance of accelerated first-order optimization algorithms in the presence of additive white stochastic disturbances. 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, as the condition number increases, variance amplification of both Nesterov's accelerated method and the heavy-ball method by Polyak is significantly larger than that of the standard gradient descent. In the context of distributed computation over networks, we examine the role of network topology and spatial dimension on the performance of these first-order algorithms. For d-dimensional tori, we establish explicit asymptotic dependence for the variance amplification on the network size and the corresponding condition number. Our results demonstrate detrimental influence of acceleration on amplification of stochastic disturbances and suggest that metrics other than convergence rate have to be considered when evaluating performance of optimization algorithms.**

Index Terms— **Consensus networks, convex optimization, first-order methods, gradient descent, heavy-ball method, Nesterov's accelerated gradient method, stochastic performance, variance amplification.**

I. Introduction

First-order methods are the workhorse for solving many large scale optimization problems [1]–[3]. While gradient descent and stochastic gradient descent are widely used because of their simplicity and robust performance, the accelerated first-order methods gained popularity because of their optimal convergence rate [4]–[9]. The behavior of these algorithms has been extensively studied in the literature under different step size selection rules [10].

In many applications, the exact value of the objective function and/or of its gradient is not available to the optimizer. This happens when the objective function is obtained via costly simulations (e.g., tuning of hyper-parameters in supervised/unsupervised learning [11]–[13]) or when the evaluation of the objective function is based on noisy measurements (e.g., real-time and embedded applications). Another application arises in the (batch) stochastic gradient settings where at each iteration the gradient of the objective function is computed from a small batch of data points. Such a batch gradient is known to be a noisy unbiased estimator of the training loss. Moreover, noise may be added to the gradient to improve generalization or to escape saddle points [14], [15].

In all of the above scenarios, only noisy unbiased estimates of the gradient of the objective function are available to the iterative algorithms. This motivates the analysis of performance of the gradient descent and its accelerated variants in the presence of noisy/inexact gradient oracle [16]–[19]. For example, while early stochastic approximation results suggest the use of step size that is inversely proportional to the iteration number [17], a more robust behavior can be obtained by combining larger step sizes with averaging [18], [20]. Utility of these averaging schemes and their modifications for solving quadratic optimization and manifold problems has also been studied [21]–[23]. Furthermore, it has been shown that accelerated first-order algorithms can tolerate less noise than their non-accelerated variants [19], [24]–[28]. In particular, an upper bound on the effect of errors on iterates of inexact accelerated proximal gradient methods was established in [26]. Using this upper bound, the authors of [26] showed that both proximal gradient and accelerated proximal gradient can maintain their convergence rate provided that the error vanishes fast enough. While the effect of imperfections on the performance of these algorithms has been extensively studied, the influence of acceleration on noise amplification has not been precisely characterized. In particular, the existing results neither pay attention to how *the distribution of the eigenvalues of the Hessian influences noise amplification* nor *provide any lower bounds on the variance amplification*.

As a first step towards answering such question, we investigate performance of first-order optimization algorithms in the presence of additive stochastic disturbances. More specifically, we focus on convex quadratic optimization problems and compare the steady-state variance amplification of accelerated methods to that of the gradient descent. For this class of problems, first-order algorithms can be cast as linear dynamical systems and tools from control theory can be used to study fundamental performance limitations and trade offs that acceleration introduces. In particular, we utilize the H_2 analysis to demonstrate poor performance of accelerated methods compared to the standard gradient descent in the presence of noise.

For unconstrained strongly convex quadratic problems, we evaluate the steady-state variance of the optimization variable for three first-order algorithms: the gradient descent, Nesterov's accelerated method and the heavy-ball method by

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

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

Polyak. We develop explicit formulae for the steady-state variance amplification of these algorithms in terms of the algorithmic parameters and problem data. For a model of a distributed computation over networks, we specialize the aforementioned asymptotic variance relations by focusing on the consensus problem for large networks with tori topologies. For such networks, we demonstrate that while gradient descent shows similar stochastic performances as that of the standard consensus algorithm [29], for large torus interconnects of spatial dimension less than five, accelerated methods suffer from significantly larger steady-state variance amplification the gradient descent.

II. Problem formulation

The unconstrained optimization problem

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

can be solved in many different ways when $f\colon \mathbb{R}^n \to \mathbb{R}$ is a convex function with an L_f-Lipschitz continuous gradient. We are interested in examining properties of three popular first-order algorithms for solving problem (1) in the presence of additive stochastic disturbances. In particular, we consider noisy versions of gradient decent,

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

Nesterov's accelerated method,

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

and Polyak's heavy-ball method,

$$x^{k+2} \;=\; x^{k+1} \;+\; \beta^k (x^{k+1} - x^k) \;-\; \alpha^k \nabla f(x^{k+1}) \;+\; w^k$$

where w^k is a stochastic disturbance, α^k is the step size, and β^k is a scalar parameter.

In the absence of strong convexity, Nesterov's accelerated method enjoys an optimal convergence rate among all linear first-order algorithms. While the gradient descent with the step size $1/L_f$ provides $(1/k)$ decay rate,

$$f(x^k) \;-\; f(x^\star) \;\leq\; \frac{L_f}{2k} \, \|x^0 \;-\; x^\star\|^2$$

accelerated Nesterov's algorithm with $\alpha = 1/L_f$ and $\beta^k = (k-1)/(k+2)$ yields,

$$f(x^k) \;-\; f(x^\star) \;\leq\; \frac{4L_f}{(k \;+\; 2)^2} \, \|x^0 \;-\; x^\star\|^2.$$

On the other hand, the heavy-ball method may even fail to converge for general strongly convex functions [30].

A. Strongly convex quadratic problems

To obtain analytical insight, in this paper we quantify the variance amplification of noisy first-order algorithms for strongly convex quadratic optimization problems. For this class of problems, the function f in (1) is given by

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

Method	Parameter choice	Rate
Gradient	$\alpha = \frac{2}{L_f + m_f}$	$\frac{\kappa - 1}{\kappa + 1}$
Nesterov	$\alpha = \frac{4}{3L_f + m_f}, \qquad \beta = \frac{\sqrt{3\kappa+1}-2}{\sqrt{3\kappa+1}+2}$	$\frac{\sqrt{3\kappa+1}-2}{\sqrt{3\kappa+1}}$
Polyak	$\alpha = \frac{4}{(\sqrt{L_f}+\sqrt{m_f})^2}, \beta = \frac{(\sqrt{\kappa}-1)^2}{(\sqrt{\kappa}+1)^2}$	$\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$

TABLE I: Optimal parameters and the exponential convergence rate bounds for m_f-strongly convex quadratic objective functions with L_f-Lipschitz gradients and $\kappa := L_f/m_f$.

where $Q = Q^T \succ 0$ is a positive definite matrix and r is a vector. In this case, gradient descent, Nesterov's accelerated method, and Polyak's heavy-ball algorithm converge exponentially to the optimal solution $x^\star = Q^{-1} r$. Furthermore, the constant values of α and β given in Table I provide optimal decay rates [30].

In what follows, without loss of generality, we set $r = 0$. This yields $\nabla f(x) = Qx$ and the above first-order algorithms can be described by a linear time-invariant state-space model,

$$\begin{aligned} \psi^{k+1} \;&=\; A \, \psi^k \;+\; B \, w^k \\ y^k \;&=\; C \, \psi^k \end{aligned} \tag{2}$$

with input w^k, output $y^k := x^k$, and state ψ^k. In particular, for gradient descent we can choose

$$\psi^k \;:=\; x^k, \;\; A \;=\; I \;-\; \alpha \, Q, \;\; B \;=\; C \;=\; I$$

where I is the identity matrix. Similarly, for Nesterov's and Polyak's methods, if we let

$$\psi^k \;:=\; \left[\begin{array}{c} x^k \\ x^{k+1} \end{array} \right], \;\; B \;=\; \left[\begin{array}{c} 0 \\ I \end{array} \right], \;\; C \;=\; \left[\begin{array}{cc} I & 0 \end{array} \right]$$

the corresponding A-matrices are, respectively, given by

$$A \;=\; \left[\begin{array}{cc} 0 & I \\ -\beta(I - \alpha Q) & (1+\beta)(I - \alpha Q) \end{array} \right]$$

$$A \;=\; \left[\begin{array}{cc} 0 & I \\ -\beta I & (1+\beta) I - \alpha Q \end{array} \right].$$

It can be shown that for α and β in Table I and $Q \succ 0$, the eigenvalues of the matrix A are inside the open unit disk in the complex plane, thereby implying stability of (2) for all three cases.

B. Variance amplification

We are interested in studying the performance of the above first-order algorithms in the presence of noise. For white stochastic disturbances w^k with

$$\mathbb{E}\left(w^k\right) \;=\; 0, \quad \mathbb{E}\left(w^k (w^l)^T\right) \;=\; I \delta(k-l)$$

we evaluate the steady-state variance of the output y^k in (2),

$$J \;:=\; \lim_{k \to \infty} \; \mathbb{E}\left(\|y^k\|^2\right)$$

where $\mathbb{E}$ is the expected value and δ is the Kronecker delta. Since (2) is stable, this quantity is well-defined and it can be equivalently expressed in terms of the steady-state covariance matrix $Y := \lim_{k\to\infty} \mathbb{E}\,(y^k(y^k)^T)$ of the output y^k,

$$J \;=\; \mathrm{trace}\,(Y)\,.$$

It is well-known that the state covariance matrix $P^k := \mathbb{E}\,(\psi^k(\psi^k)^T)$ satisfies the recursive Lyapunov equation

$$P^{k+1} \;=\; A\,P^k A^T \;+\; BB^T$$

and that the steady-state limit $P := \lim_{k\to\infty} P^k$ can be obtained from the algebraic Lyapunov equation,

$$P \;=\; A\,PA^T \;+\; BB^T. \tag{3}$$

The solution to this equation and the fact that $Y = CPC^T$ allow us to compute the variance amplification as

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

Finally, we note that this measure of the input-output amplification determines the square of the H_2 norm of system (2). It is well-known that the H_2 norm also has an appealing deterministic interpretation in terms of the L_2 norm of the system's impulse response [31].

III. Analytical expressions

In this section, we provide analytical expressions for the variance amplification of first-order algorithms for strongly convex quadratic optimization problems. We show that, in addition to algorithmic parameters α and β, the variance amplification depends on all eigenvalues of the positive definite matrix Q. This should be compared and contrasted to the optimal rate of linear convergence which only depends on the ratio of the largest and smallest eigenvalues of Q.

A. Influence of the eigenvalues on variance amplification

We use the eigenvalue decomposition of $Q = V\Lambda V^T \succ 0$, where $\Lambda = \mathrm{diag}\,(\lambda_i)$ is a diagonal matrix of the eigenvalues and V is a unitary matrix of the eigenvectors of Q to bring A, B, and C in (2) into a block diagonal form,

$$\hat{A} \;=\; \mathrm{diag}\,(\hat{A}_i),\; \hat{B} \;=\; \mathrm{diag}\,(\hat{B}_i),\; \hat{C} \;=\; \mathrm{diag}\,(\hat{C}_i),\quad i \;=\; 1,\ldots,n.$$

In particular, the unitary coordinate transformation

$$\hat{x}^k \;:=\; V^T x^k,\;\; \hat{w}^k \;:=\; V^T w^k \tag{4}$$

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

$$\hat{\psi}_i^k \;=\; \hat{x}_i^k,\; \hat{A}_i \;=\; 1 \;-\; \alpha\lambda_i,\; \hat{B}_i \;=\; \hat{C}_i \;=\; 1.$$

Similarly, for accelerated algorithms, change of coordinates (4) in conjunction with a permutation of variables yield

$$\hat{\psi}_i^k \;=\; \begin{bmatrix} \hat{x}_i^k \\ \hat{x}_i^{k+1} \end{bmatrix},\; \hat{B}_i \;=\; \begin{bmatrix} 0 \\ 1 \end{bmatrix},\; \hat{C}_i \;=\; \begin{bmatrix} 1 & 0 \end{bmatrix}$$

where

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

for Nesterov's accelerated method, and

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

for the heavy-ball method.

This block diagonal structure allows us to explicitly solve Lyapunov equation (3) for P and derive an analytical expression for the variance amplification in terms of the eigenvalues λ_i of the matrix Q and the algorithmic parameters α and β. Namely, under coordinate transformation (4) and a suitable permutation of variables, equation (3) can be brought into an equivalent set of algebraic Lyapunov equations,

$$\hat{P}_i \;=\; \hat{A}_i\,\hat{P}_i\,\hat{A}_i^T \;+\; \hat{B}_i\hat{B}_i^T,\quad i \;=\; 1,\ldots,n \tag{5}$$

where $\hat{P}_i$ is a scalar for the gradient descent method and a 2×2 matrix for the accelerated algorithms. In Theorem 1, for each of these algorithms, we determine an analytical solution $\hat{P}_i$ to (5) and express the output variance as

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

Theorem 1: For strongly convex quadratic problems, the variance amplification of first-order algorithms subject to additive white stochastic disturbances with zero mean and unit variance is given by $J = \sum_{i=1}^n \hat{J}(\lambda_i)$, where

$$\begin{aligned}
\text{Gradient:}\quad & \hat{J}(\lambda) \;=\; \tfrac{1}{\alpha\lambda(2-\alpha\lambda)} \\
\text{Nesterov:}\quad & \hat{J}(\lambda) \;=\; \tfrac{1+\beta(1-\alpha\lambda)}{\alpha\lambda(1-\beta(1-\alpha\lambda))(2(1+\beta)-(2\beta+1)\alpha\lambda)} \\
\text{Polyak:}\quad & \hat{J}(\lambda) \;=\; \tfrac{1+\beta}{1-\beta}\tfrac{1}{\alpha\lambda(2(1+\beta)-\alpha\lambda)}.
\end{aligned}$$

Proof: For gradient descent, $\hat{A}_i = 1-\alpha\lambda_i$ and $\hat{B}_i = 1$ are both scalars and, hence, the solution to (5) is given by

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

For 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 (5) is

$$\hat{P}_i \;=\; \begin{bmatrix} p_i & b_ip_i/(1-a_i) \\ b_ip_i/(1-a_i) & p_i \end{bmatrix}$$

where

$$p_i \;:=\; \frac{a_i\;-\;1}{(a_i\;+\;1)(b_i\;+\;a_i\;-\;1)(b_i\;-\;a_i\;+\;1)}. \tag{6a}$$

For Nesterov's accelerated method, we have

$$a_i \;=\; -\beta(1\;-\;\alpha\lambda_i),\; b_i \;=\; (1\;+\;\beta)(1\;-\;\alpha\lambda_i) \tag{6b}$$

and for the heavy-ball method,

$$a_i \;=\; -\beta,\; b_i \;=\; (1\;+\;\beta)\;-\;\alpha\lambda_i. \tag{6c}$$

Now, since $\hat{C}_i = 1$ for gradient descent and $\hat{C}_i = [\,1\;\;0\,]$ for

accelerated algorithms, we have

$$\hat{J}(\lambda_i) \;:=\; \mathrm{trace}\,(\hat{C}_i\,\hat{P}_i\,\hat{C}_i^T) \;=\; p_i.$$

Finally, if we use the expression for p_i for gradient descent and substitute for a_i and b_i in (6a) from (6b) for Nesterov's algorithm and from (6c) for Polyak's algorithm we obtain the expressions for $\hat{J}(\lambda_i)$ in the statement of the theorem. ■

B. Analysis in condition number

For the optimal parameters provided in Table I, we use the expressions derived in Theorem 1 to establish explicit relations between the variance amplification of gradient descent and its accelerated variants. We also provide tight upper and lower bounds that reveal the dependence on the condition number κ and the problem size n.

In what follows, we use $J_{\rm gd}$, $J_{\rm na}$, and $J_{\rm hb}$ to denote the steady-state variances of gradient descent, Nesterov's accelerated, and Polyak's heavy-ball methods. Theorem 2 establishes a linear relation between $J_{\rm hb}$ and $J_{\rm gd}$ and demonstrates that acceleration always increases variance amplification of the heavy-ball method relative to gradient descent.

Theorem 2: For strongly convex quadratic problems with $\lambda_{\max}(Q) = L_f$, $\lambda_{\min}(Q) = m_f$, $\kappa := L_f/m_f$, and parameters given in Table I, the variance amplification of the heavy-ball method is determined by

$$J_{\rm hb} \;=\; \frac{(\sqrt{\kappa}+1)^4}{8\sqrt{\kappa}(\kappa+1)}\,J_{\rm gd}.$$

Proof: For the values of the parameters α and β provided in Table I, using Theorem 1, it is straightforward to show that for any $\lambda \in [m_f, L_f]$ we have

$$\frac{\hat{J}_{\rm hb}(\lambda)}{\hat{J}_{\rm gd}(\lambda)} \;=\; \frac{(\sqrt{\kappa}+1)^4}{8\sqrt{\kappa}(\kappa+1)}$$

which proves the result. ■

Theorem 2 shows that the ratio between $J_{\rm hb}$ and $J_{\rm gd}$ is a monotonically increasing function of the condition number κ. We note that although the heavy-ball method decreases the number of iterations required to achieve any given accuracy by a factor of $\sqrt{\kappa}$, it also increases variance amplification relative to gradient descent by a factor that asymptotically (i.e., as $\kappa \to \infty$) scales as $\sqrt{\kappa}$. The relationship between variance amplification of Nesterov's method and gradient descent is more subtle and it depends on the distribution of the eigenvalues of the matrix Q.

In the next proposition, we provide upper and lower bounds on variance amplification in terms of the problem size n and the condition number κ.

Proposition 1: For strongly convex quadratic problems with $\lambda_{\max}(Q) = L_f$, $\lambda_{\min}(Q) = m_f$, and $\kappa := L_f/m_f$, the variance amplification of noisy first-order algorithms, with parameters given in Table I, is bounded by

$$\frac{(\kappa-1)^2}{2\kappa} + n \;\leq\; J_{\rm gd} \;\leq\; \frac{n(\kappa+1)^2}{4\kappa}$$

$$\frac{3\kappa\sqrt{3\kappa}}{32} + n - 1 \;\leq\; J_{\rm na} \;\leq\; \frac{6n(\kappa+3)\sqrt{\kappa+3}}{32}$$

$$\frac{(\kappa-1)^2}{2\kappa} + n \;\leq\; J_{\rm hb}\,\frac{8\sqrt{\kappa}(\kappa+1)}{(\sqrt{\kappa}+1)^4} \;\leq\; \frac{n(\kappa+1)^2}{4\kappa}.$$

Proof: It is straightforward to show that for the values of parameters α and β given in Table I, the functions $\hat{J}(\lambda)$ in Theorem 1 for gradient descent and Nesterov's algorithm attain their maximum and minimum at $\lambda = m_f$ and $\lambda = 1/\alpha$, respectively. In other words,

$$1 \;=\; \hat{J}(1/\alpha) \;\leq\; \hat{J}(\lambda) \;\leq\; \hat{J}(m_f), \quad \forall \lambda \in [m_f, L_f]. \quad (7)$$

Therefore, fixing the smallest and largest eigenvalues, the variance amplification J is maximized when the other $n-2$ eigenvalues are all equal to m_f and is minimized when they are all equal to $1/\alpha$. This leads to the above upper and lower bounds for gradient descent. For Nesterov's algorithm, it is not hard to verify that

$$\frac{3\kappa\sqrt{3\kappa}}{32} \;\leq\; \hat{J}_{\rm na}(m_f) \;\leq\; \frac{6(\kappa+3)\sqrt{\kappa+3}}{32}.$$

Combining this inequality with (7) completes the proof for Nesterov's algorithm. The bounds for the heavy-ball method is the direct consequence of applying Theorem 2 to the bounds obtained for gradient descent method. ■

For a fixed problem size n, we observe that the variance amplification of gradient descent $J_{\rm gd}$ scales linearly with the condition number. On the other hand, the variance amplification of the accelerated algorithms scales as $\kappa\sqrt{\kappa}$. However, in many problems, the condition number κ depends on the problem size n which calls for a closer examination of the scaling trends as n increases. This issue is further discussed in Section IV. More specifically, in the next section, we consider a simple model for a distributed computation over undirected consensus networks and employ the first-order algorithms to compute the network average. We utilize the relations in Theorems 1 and 2 to establish bounds on the variance amplification of noisy algorithms as the size of the network increases.

IV. Application to distributed computation over network

Distributed computation over networks has been the focus of extensive research in the optimization and machine learning communities. In this problem, the goal is to optimize an objective function (e.g., for the purpose of training a model) over distributed processing units that are connected to each other over a network. Clearly, the structure of the network (e.g., node dynamics and network topology) would impact the performance (e.g., noise amplification) of any optimization algorithm.

As a first step toward understanding the impact of network structure on noisy optimization algorithms, in this section,

we analyze the asymptotic noise amplification in solving standard distributed consensus problem via first-order algorithms. In the classical network consensus problem, the sensors/processing nodes each have measurements and the goal is to compute the average of these measurements by local information exchange among the nodes. This problem arises in a number of applications ranging from social networks, to distributed computing networks, to cooperative control in multi-agent systems and it can be formulated as a minimization of the quadratic function

$$\underset{x}{\text{minimize}} \ \frac{1}{2}\, x^T L\, x \tag{8}$$

where $L = L^T \succeq 0$ is the Laplacian matrix of the graph associated with the underlying undirected network and $x \in \mathbb{R}^n$ is the vector of node measurements. The graph Laplacian of a connected network is a positive semidefinite matrix that has only one zero eigenvalue with the corresponding eigenvector of all ones, $\mathbb{1} := [\,1 \ \cdots \ 1\,]^T$. Using this fact, one can make the following observations:

- Any vector parallel to the vector of all ones $\mathbb{1}$ is a global minimizer of (8).
- In the absence of noise, each step of the gradient descent or its accelerated variants does not change the average values of nodes, i.e., $\mathbb{1}^T x$ remains constant during the algorithm.
- Performing each step of gradient descent, Nesterov's method, or the heavy-ball method only requires local information exchange among the nodes.

Combining these observations, one can conclude that, in the absence of noise, all three algorithms converge to the average of the node values, $\frac{1}{n}\sum_i x_i \mathbb{1}$. Moreover, these algorithms can be implemented in a completely distributed fashion. In the presence of noise, the zero eigenvalue of the graph Laplacian leads to an unbounded steady-state variance of x^k because the network average $\frac{1}{n}\sum_i x_i$ experiences random walk. As described in [32], performance of noisy algorithms can be quantified by examining the mean-square deviation of the node values from the network average. We note that, although (8) is not strongly convex on $\mathbb{R}^n$, for connected undirected networks it is strongly convex on the orthogonal complement of the subspace spanned by the vector of all ones, $\mathbb{1}^\perp$. On this subspace, only positive eigenvalues of L influence the variance amplification and all formulae in Theorems 1 and 2 all remain valid by conducting summations over the non-zero eigenvalues of L.

In what follows, we study large networks with simple structure for which there is a systematic way of increasing the network size n. When the computation at each node is subject to additive noise we evaluate asymptotic performance of the noisy first-order algorithms as the size of the network increases. We show that the asymptotic variance amplification of the gradient descent is equivalent to that of the standard consensus algorithm studied in [29] and that acceleration negatively impacts the performance of noisy algorithms. Our results also highlight the subtle influence of the distribution of the eigenvalues on variance amplification.

A. Explicit formulae for d-dimensional tori networks

Tori with nearest neighbor interactions generalize one-dimensional rings to higher spatial dimensions. Let $\mathbb{Z}_m$ denote the group of integers modulo m. A d-dimensional torus $\mathbb{T}_m^d$ consists of $n := m^d$ nodes denoted by v_a where $a \in \mathbb{Z}_m^d$ and a set of edges $E := \{\{v_a \ v_b\} \,|\, \|a - b\| = 1 \mod m\}$; nodes v_a and v_b are neighbors if and only if a and b differ exactly at one entry by one. For example, $\mathbb{T}_m^1$ denotes a ring with $n = m$ nodes and $\mathbb{T}_m^5$ denotes a five dimensional torus with $n = m^5$ nodes.

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

$$\lambda_i \ = \ \sum_{l\,=\,1}^{d} 2\left(1 \,-\, \cos\tfrac{2\pi i_l}{m}\right), \quad i_l \,\in\, \mathbb{Z}_m \tag{9}$$

where $i := (i_1, \ldots, i_d) \in \mathbb{Z}_m^d$. We note that $\lambda_0 = 0$ is the only zero eigenvalue of the graph Laplacian with eigenvector of all ones and that all other eigenvalues are positive. As aforementioned, all formulae in Theorem 1 remain valid if sums are taken over non-zero eigenvalues of L. Namely, the network-size normalized variance amplification J/n of noisy algorithms can be written as

$$\frac{J}{n} \ = \ \frac{1}{n} \sum_{0 \,\neq\, i \,\in\, \mathbb{Z}_m^d} \hat{J}(\lambda_i)$$

where, for each algorithm, $\hat{J}\colon \mathbb{R}_+ \to \mathbb{R}_+$ is a function determined in Theorem 1 and λ_i's are the non-zero eigenvalues of the graph Laplacian L.

The following theorem characterizes the asymptotic variance amplification per network node of noisy first-order distributed averaging algorithms for d-dimensional tori networks with $n = m^d$ nodes. This result is obtained by combining relations derived in Section III with the analytical expression (9) for the eigenvalues of the corresponding Laplacian matrix $L \in \mathbb{R}^{n\times n}$. The key observation is that, for $m \gg 1$, the condition number of L on $\mathbb{1}^\perp$ is given by

$$\kappa \ = \ \frac{\lambda_{\max}}{\lambda_{\min}} \ \approx \ \frac{2}{1 \,-\, \cos\frac{2\pi}{m}} \ = \ \Theta(m^2) \ = \ \Theta(n^{2/d}).$$

Here, the notation $g = \Theta(h)$ means that there exist positive constants c_1 and c_2 such that

$$c_1 \ \leq \ \liminf_{h\,\to\,\infty} \frac{h}{g} \ \leq \ \limsup_{h\,\to\,\infty} \frac{h}{g} \ \leq \ c_2.$$

Theorem 3: Let $L \in \mathbb{R}^{n\times n}$ be the graph Laplacian of the d-dimensional undirected torus $\mathbb{T}_m^d$ with $n = m^d \gg 1$ nodes. For optimization problem (8), the variance amplification J of the noisy first-order algorithms on the subspace $\mathbb{1}^\perp$ satisfy the asymptotic trends provided in Table II.

Proof: The proof is omitted due to page limitations. ■

It is worth noting that for one-dimensional rings, the upper bounds on the variance amplification established in Proposition 1 are all conservative by a factor of $\sqrt{\kappa}$ when compared to the exact trends presented in Theorem 3 that are obtained based on the distribution of the eigenvalues. This

	$d=1$	$d=2$	$d=3$	$d=4$
$\frac{J_{\rm gd}}{n}$	$\Theta(\sqrt{\kappa})$	$\Theta(\log\,\kappa)$	$\Theta(1)$	$\Theta(1)$
$\frac{J_{\rm na}}{n}$	$\Theta(\kappa)$	$\Theta(\sqrt{\kappa}\log\,\kappa)$	$\Theta(\kappa^{\frac{1}{4}})$	$\Theta(\log\,\kappa)$
$\frac{J_{\rm hb}}{n}$	$\Theta(\kappa)$	$\Theta(\sqrt{\kappa}\,\log\,\kappa)$	$\Theta(\sqrt{\kappa})$	$\Theta(\sqrt{\kappa})$

TABLE II: The asymptotic trends for the network-size-normalized variance amplification J/n of the first-order methods for the torus $\mathbb{T}_m^d$ with $n = m^d$ nodes. Here, $\kappa = \Theta(n^{2/d})$ is the condition number of the graph Laplacian L restricted to the subspace $\mathbb{1}^\perp$.

gap becomes even larger in higher spatial dimensions mainly due to the fact that the upper bounds are obtained using only the extreme eigenvalues. In other words, we observe that as we increase the dimension, the trends depart from the upper bounds and get closer to the lower bounds established in Proposition 1. Our results emphasize the importance of paying attention to the whole spectrum of the Hessian rather than just the extreme eigenvalues.

V. Concluding remarks

We demonstrate that acceleration techniques although can improve the convergence rate of first-order algorithms, may suffer from a significantly larger noise amplfication. In particular, in the presence of white additive stochastic disturbances, as we establish in Section III-B, for large condition number κ and fixed problem size n, gradient descent has a smaller steady-state variance for the optimization variable than both Nesterov's accelerated and the heavy-ball methods.

We further consider the special case of quadratic optimization problem whose objective function is given by the Laplacian matrix of a torus interconnect that is a generalization of cycle graphs. For this class of problems, that arise in the context of distributed computation over networks, we derive asymptotic relations for the variance amplification of the three above algorithms, as the size of the network increases. These results demonstrate that for tori networks gradient decent is more robust to stochastic disturbances than its accelerated variants.

References

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

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

[3] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *arXiv preprint arXiv:1606.04838*, 2016.

[4] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *International Conference on Machine Learning*, 2013, pp. 1139–1147.

[5] B. T. Polyak, "Some methods of speeding up the convergence of iteration methods," *USSR Computational Mathematics and Mathematical Physics*, vol. 4, no. 5, pp. 1–17, 1964.

[6] Y. Nesterov, "A method of solving a convex programming problem with convergence rate o (1/k2)," in *Soviet Mathematics Doklady*, vol. 27, 1983, pp. 372–376.

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

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

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

[10] M. Fazlyab, A. Ribeiro, M.Morari, and V. M. Preciado, "Analysis of optimization algorithms via integral quadratic constraints: Nonstrongly convex problems," *arXiv preprint arXiv:1705.03615v2*, 2018.

[11] D. Maclaurin, D. Duvenaud, and R. Adams, "Gradient-based hyperparameter optimization through reversible learning," in *International Conference on Machine Learning*, 2015, pp. 2113–2122.

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

[13] A. Beirami, M. Razaviyayn, S. Shahrampour, and V. Tarokh, "On optimal generalizability in parametric learning," in *Advances in Neural Information Processing Systems*, 2017, pp. 3458–3468.

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

[15] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan, "How to escape saddle points efficiently," *arXiv preprint arXiv:1703.00887*, 2017.

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

[17] H. Robbins and S. Monro, "A stochastic approximation method," *The Annals of Mathematical Statistics*, pp. 400–407, 1951.

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

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

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

[21] A. Dieuleveut, N. Flammarion, and F. Bach, "Harder, better, faster, stronger convergence rates for least-squares regression," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 3520–3570, 2017.

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

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

[24] M. Baes, "Estimate sequence methods: extensions and approximations," *Institute for Operations Research, ETH, Zürich, Switzerland*, 2009.

[25] A. d'Aspremont, "Smooth optimization with approximate gradient," *SIAM Journal on Optimization*, vol. 19, no. 3, pp. 1171–1183, 2008.

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

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

[28] S. Cyrus, B. Hu, B. V. Scoy, and L. Lessard, "A robust accelerated optimization algorithm for strongly convex functions," *arXiv preprint arXiv:1710.04753v2*, 2018.

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

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

[31] G. E. Dullerud and F. Paganini, *A course in robust control theory: a convex approach*. New York: Springer-Verlag, 2000.

[32] 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.