SIAM J. OPTIM.
Vol. 30, No. 1, pp. 717–751
© 2020 Society for Industrial and Applied Mathematics

Downloaded 05/03/20 to 18.9.61.111. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

ROBUST ACCELERATED GRADIENT METHODS FOR SMOOTH STRONGLY CONVEX FUNCTIONS*

NECDET SERHAT AYBAT†, ALIREZA FALLAH‡, MERT GÜRBÜZBALABAN§, AND ASUMAN OZDAGLAR‡

Abstract. We study the trade-offs between convergence rate and robustness to gradient errors in designing a first-order algorithm. We focus on gradient descent and accelerated gradient (AG) methods for minimizing strongly convex functions when the gradient has random errors in the form of additive white noise. With gradient errors, the function values of the iterates need not converge to the optimal value; hence, we define the robustness of an algorithm to noise as the asymptotic expected suboptimality of the iterate sequence to input noise power. For this robustness measure, we provide exact expressions for the quadratic case using tools from robust control theory and tight upper bounds for the smooth strongly convex case using Lyapunov functions certified through matrix inequalities. We use these characterizations within an optimization problem which selects parameters of each algorithm to achieve a particular trade-off between rate and robustness. Our results show that AG can achieve acceleration while being more robust to random gradient errors. This behavior is quite different than previously reported in the deterministic gradient noise setting. We also establish some connections between the robustness of an algorithm and how quickly it can converge back to the optimal solution if it is perturbed from the optimal point with deterministic noise. Our framework also leads to practical algorithms that can perform better than other state-of-the-art methods in the presence of random gradient noise.

Key words. convex optimization, stochastic approximation, robust control theory, accelerated methods, Nesterov's method, matrix inequalities

AMS subject classifications. 62L20, 90C15, 90C25, 90C30, 93C05, 93C10

DOI. 10.1137/19M1244925

1. Introduction.

For many large-scale convex optimization and machine learning problems, first-order methods have been the leading computational approach for computing low- to medium-accuracy solutions because of their cheap iterations and mild dependence on the problem dimension and data size. The typical analysis of first-order methods assumes the availability of exact gradient information and provides statements on the rate of convergence to the optimal solution as the main performance criterion. However, in many applications, the gradient contains deterministic or stochastic errors either because the gradient is computed by inexactly solving an auxiliary problem [10, 13], or the method itself involves errors with respect to the full gradient as in standard incremental gradient, stochastic gradient, and stochastic approximation methods [4, 5, 39, 41]. When there are persistent errors in gradients, the iterates do not converge and could oscillate in a neighborhood of the optimal solution or may even diverge [4, 5, 13, 19]. This makes robustness of the algorithms

*Received by the editors February 14, 2019; accepted for publication (in revised form) October 29, 2019; published electronically February 27, 2020. The authors are listed in alphabetical order.
https://doi.org/10.1137/19M1244925

Funding: The work of the first author is partially supported by NSF grant CMMI-1635106. The second author is partially supported by a Siebel Scholarship, NSF grants DMS-1723085 and CCF-1814888. The third author receives support from NSF grants DMS-1723085 and CCF-1814888.

†Department of Industrial and Manufacturing Engineering, Pennsylvania State University, University Park, PA 16802 (nsa10@psu.edu).

‡Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139 (afallah@mit.edu, asuman@mit.edu).

§Department of Management Science and Information Systems, Rutgers University, Piscataway, NJ 08854 (mg1366@rutgers.edu).

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 05/03/20 to 18.9.61.111. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

to gradient errors (in terms of solution accuracy) another important performance objective [13, 27]. In particular, even though the accelerated gradient (AG) method proposed by Nesterov converges faster than gradient descent (GD) in the absence of noise for convex problems [36], it was shown that AG methods are less robust to errors, i.e., accelerated methods require higher precision gradient information than GD to achieve the same solution accuracy [10, 13, 19, 43].

In this paper, we study the trade-offs between convergence rate and robustness to gradient errors in designing a first-order algorithm. We focus on GD and Nesterov's AG method for minimizing strongly convex smooth functions when the gradient has stochastic errors and investigate how the parameters of each algorithm should be set to achieve a particular trade-off between these two performance objectives. To study this question systematically, we employ tools from control theory whereby we represent each of the algorithms as a dynamical system. This approach has attracted recent attention and has already led to a number of insights for the design and analysis of optimization algorithms [9, 18, 28, 29, 33, 45]. The novelty of our work is to use this approach to provide explicit characterizations of robustness, which can then be placed in a computationally tractable optimization problem for selecting the algorithm parameters to systematically achieve a desired trade-off.

We first focus on problems with a strongly convex quadratic objective function. For this case, the rate of convergence of any of the two algorithms we study is given by the spectral radius of the "state-transition" matrix in the dynamical system representation. To characterize robustness, we consider the asymptotic expected suboptimality for the centered iterate sequence (output vector of the dynamical system) per unit noise which is a measure of the asymptotic accuracy of the iterates. For the quadratic case we show that this limit exists and can be characterized using the H_2 *norm* of a transformed linear dynamical system. The H_2 norm is a fundamental measure for quantifying robustness of a linear system to noise [46] and admits various definitions and characterizations. We focus on a particular representation of the H_2 norm that requires the solution of a discrete Lyapunov equation. This representation leads to explicit expressions for robustness of GD and AG.

Using this result, we study the rate and robustness trade-off of the GD method for minimizing quadratic strongly convex functions. The spectral radius of the state-transition matrix corresponding to GD dynamics, hence, the rate of convergence for GD, can be expressed in terms of the smallest and largest eigenvalues of the positive definite matrix Q defining the Hessian of the strongly convex quadratic objective. We show that our robustness measure admits a tractable characterization for GD in terms of the spectrum of Q. We also show a fundamental lower bound on the robustness level of an algorithm for any achievable convergence rate.

We next consider the AG method defined by two parameters: stepsize α and momentum parameter β. Our first step is to characterize the *stability region* of the method, i.e., the set of nonnegative (α, β) for which the spectral radius of the state-transition matrix is less than or equal to one. Similar to GD, we then provide an explicit characterization of the H_2 norm of the dynamical system representation of AG. We use these explicit expressions for both GD and AG within an optimization problem for selecting the parameters to minimize the robustness measure subject to a given upper bound on the convergence rate. Our results show that AG with properly selected parameters is superior to GD in the sense that AG can achieve the same rate with GD while being more robust to noise; similarly, AG can be tuned to be faster than GD while achieving the same robustness level. This behavior contrasts with the comparison of GD and AG in the deterministic gradient error setting in [13], which

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 05/03/20 to 18.9.61.111. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

shows GD performance degrades gracefully while AG may accumulate error. These results show the random and deterministic noise settings have different behavior.

In our second set of results, we extend our analysis to handle minimization of strongly convex smooth functions, i.e., $\min_{x\in\mathbb{R}^d} f(x)$. In this setting, the dynamical system representation of a first-order algorithm will no longer be a linear system due to the nonlinear gradient map, ∇f. The analysis in this section is not limited to GD or AG; in particular, given a first-order optimization algorithm, we use a linear dynamic system with *nonlinear* feedback to model the dynamic behavior of the algorithm. For these systems, we again use a robustness measure that can be seen as a discrete-time version of a more general H_2 norm for nonlinear systems [20]—see also [44] for a similar definition given for linear systems with nonlinear feedback. We derive upper bounds on the robustness measures for GD and AG using Lyapunov functions certified through matrix inequalities (MIs) and investigate the trade-off between rate and robustness.

In addition to the above cited papers, Devolder's Ph.D. thesis [11] is closely related to our paper. Chapters 4 and 6 of this thesis consider smooth weakly convex functions under a deterministic oracle model, whereas Chapter 7 focus on a stochastic oracle model; these general oracles can model inexactness in the gradients as well as function evaluations. In the deterministic oracle case, Devolder shows that the primal gradient method (PGM) and the dual gradient method (DGM) on smooth weakly convex objectives exhibit slow convergence with a rate $\mathcal{O}(1/k)$ but without accumulation of errors (the total effect of errors after k iterations is equal to the individual error δ of each first-order information), whereas AG methods converge faster with rate $\mathcal{O}(\frac{1}{k^2})$ but suffer from accumulation of errors at a linear rate $\mathcal{O}(k\delta)$. Based on these observations, Devolder, Glineur, and Nesterov [12] design a novel family of first-order methods called intermediate gradient methods for solving smooth weakly convex problems; these methods have an intermediate speed and intermediate sensitivity to gradient errors, i.e., faster than classical gradient methods and more robust to noise than the AG methods. In the stochastic oracle case, Devolder developed a class of AG methods for weakly convex functions with decaying stepsize rules and showed that the expected suboptimality admits the convergence rate $\mathcal{O}(\frac{LR^2}{k^2} + \frac{\sigma R}{\sqrt{k}})$ as opposed to the $\mathcal{O}(\frac{LR^2}{k} + \frac{\sigma R}{\sqrt{k}})$ rate of PGM and DGM, where R is the distance of the initial point to the optimal solution, L is the Lipschitz constant for the gradient of the objective $f(x)$, and σ is the level of the stochastic noise [11, Chapter 7]. In his thesis, Devolder studied also smooth and strongly convex objectives under the same deterministic oracle model, showing that both PGM and DGM converge with a rate that is proportional to $\exp(-k\frac{\mu}{L})$ without accumulation of errors where μ is the strong convexity constant, whereas AG converges faster proportional to $\exp(-k\sqrt{\frac{\mu}{L}})$ while the error accumulation behaves like $\sqrt{\frac{L}{\mu}}\delta$ up to a constant [11, Chapter 5]. On the other hand, the smooth and strong convex objectives subject to stochastic errors were left as future work [11, Chapter 8.1.1], and this is the setting considered in our paper where we focus on *stochastic* additive gradient errors for *strongly convex* objectives, which arises in a number of problems in machine learning and large-scale optimization [3, 27, 40]. In this setting, Ghadimi and Lan [23, 24] propose an accelerated method called AC-SA for solving strongly convex composite optimization problems obtaining an optimal rate matching the lower complexity bounds for stochastic optimization. Flammarion and Bach [19] considered accelerated versions of GD for quadratic optimization that attain the optimal rates for both the bias and variance terms, respectively, in the performance bounds. Michalowsky and

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 05/03/20 to 18.9.61.111. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Ebenbauer [34] posed the design of deterministic gradient algorithms as a state feedback problem and used robust control theory and linear MIs to study them. Mohammadi, Razaviyayn, and Jovanović [35] examined the sensitivity of accelerated algorithms to stochastic noise for strongly convex quadratic functions in terms of the steady-state variance of the optimization variable. Finally, Dvurechensky and Gasnikov [15] consider composite convex optimization problems with inexact first-order oracles having both deterministic and stochastic errors; indeed, their inexact oracle is an extension of the one adopted in [12, 13] to include stochastic errors. For this setting, Dvurechensky and Gasnikov [15] propose a stochastic version of the intermediate gradient method in [12] and analyze the convergence rate in terms of expected suboptimality and error accumulation due to inexact oracle; the proposed algorithm in [12] has complexity bounds matching the optimal lower complexity bounds for composite convex problems with stochastic inexact oracle as in [23, 24]. Finally, Hu, Seiler, and Lessard [29] analyze the stochastic gradient method under deterministic noise and study the effect of the stepsize on the convergence rate and the asymptotic neighborhood of convergence. These papers focus on the convergence rate of the algorithms, whereas our goal is to define robustness and design algorithms to successfully trade off different objectives. Furthermore, we make some connections between the robustness of a first-order method and its behavior when perturbed from the optimal solution and show that AG is more resilient to perturbations in the sense that it recovers the optimal point with less energy compared to GD for sufficiently small stepsizes. We will also demonstrate in our numerical experiments that the framework we propose is competitive in practice with the existing state-of-the-art algorithms from the literature and can outperform them in some problems, illustrating the potential of the proposed framework in practice. In fact, in a companion paper, we use our framework to develop a universally optimal multistage stochastic gradient algorithm for stochastic optimization [1] which achieves the lower bounds without assuming a known bound for suboptimality or the variance of the gradient noise.

1.1. Preliminaries and notation. For two functions g, h defined over positive integers, we say $f = \Theta(g)$ if there exist constants C_l, C_u, and n_0 such that $C_l g(n) \leq f(n) \leq C_u g(n)$ for every positive integer $n \geq n_0$. For a set I, $|I|$ denotes the cardinality of the set I. Let $\delta[k]$ denote the Kronecker delta function, i.e., $\delta[0] = 1$ and $\delta[k] = 0$ for any integer $k \geq 1$. The $d \times d$ identity and zero matrices are denoted by I_d and 0_d, respectively. We define $\mathbf{diag}(a_1, \ldots, a_d)$ or $\mathbf{diag}([a_i]_{i=1}^d)$ as the diagonal matrix with diagonal entries $a_1, \ldots, a_d$; similarly, $\mathbf{diag}([A_i]_{i=1}^d)$ denotes a block diagonal matrix with ith block equal to $A_i \in \mathbb{R}^{n_i \times n_i}$ for $i = 1, \ldots, d$. For matrix $A \in \mathbb{R}^{d \times d}$, $\mathrm{Tr}(A)$ denotes the trace of A. We use the superscript $^\top$ to denote the transpose of a vector or a matrix depending on the context. The spectral radius of A is defined as the largest absolute value of its eigenvalues and is denoted by $\rho(A)$. We say that a square matrix A is *discrete-time stable* if all of its eigenvalues lie strictly inside the unit disc in the complex plane, i.e., if $\rho(A) < 1$. Throughout this paper, all vectors are represented as column vectors. Let $\mathbb{S}^m$ be the set of all symmetric $m \times m$ matrices. Similarly, $\mathbb{S}^m_{++}$ ($\mathbb{S}^m_+$) denotes the set of all symmetric and positive (semi)definite $m \times m$ matrices. For two matrices $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{p \times q}$, their Kronecker product is denoted by $A \otimes B$. For scalars $0 < \mu \leq L$, we define $S_{\mu,L}(\mathbb{R}^d)$ as the set of continuously differentiable functions $f : \mathbb{R}^d \to \mathbb{R}$ that are strongly convex with modulus μ and have Lipschitz-continuous gradients with constant L, i.e.,

$$\frac{L}{2}\|x - y\|^2 \geq f(x) - f(y) - \nabla f(y)^\top (x - y) \geq \frac{\mu}{2}\|x - y\|^2 \quad \forall\, x, y \in \mathbb{R}^d \tag{1.1}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 05/03/20 to 18.9.61.111. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

(see, e.g., [37]), where the gradient ∇f is represented as a column vector. The ratio $\kappa \triangleq \frac{L}{\mu}$ is called the *condition number* of f. In many places, we also use the following relation for strongly convex smooth functions.

LEMMA 1.1 (Theorem 2.1.12 in [37]). *If $f \in S_{\mu,L}(\mathbb{R}^d)$, then for every $x, y \in \mathbb{R}^d$,*

$$(\nabla f(x) - \nabla f(y))^\top (x - y) \geq \frac{\mu L}{\mu + L}\|x - y\|^2 + \frac{1}{\mu + L}\|\nabla f(x) - \nabla f(y)\|^2.$$

For our subsequent analysis, we represent the preceding relation in matrix form:

$$\begin{bmatrix} x - y \\ \nabla f(x) - \nabla f(y) \end{bmatrix}^\top \begin{bmatrix} 2\mu L I_d & -(\mu + L) I_d \\ -(\mu + L) I_d & 2 I_d \end{bmatrix} \begin{bmatrix} x - y \\ \nabla f(x) - \nabla f(y) \end{bmatrix} \leq 0 \quad \forall\, x, y \in \mathbb{R}^d. \tag{1.2}$$

2. Optimization algorithms as dynamical systems. Our goal is to design first-order algorithms with certain rate-robustness balance to solve

$$f^* \triangleq \min_{x \in \mathbb{R}^d} f(x), \quad \text{where} \quad f \in S_{\mu,L}(\mathbb{R}^d), \tag{2.1}$$

when the gradient ∇f is corrupted by random errors in the form of additive white noise. We denote the unique optimal solution of problem (2.1) by x^*. We will focus on GD and AG and show how the parameters of these algorithms can be tuned to optimize various performance metrics.

Our analysis builds on a dynamical system representation of these algorithms. A discrete-time dynamical system with a feedback rule ϕ can be expressed as

$$\xi_{k+1} = A\xi_k + Bu_k, \quad y_k = C\xi_k + Du_k, \quad u_k = \phi(y_k), \tag{2.2}$$

for $k \geq 0$, where $\xi_k \in \mathbb{R}^m$ is the *state*, $u_k \in \mathbb{R}^d$ is the *input*, and $y_k \in \mathbb{R}^d$ is the *output*. The matrices A, B, C, and D are called the *system matrices*; they are fixed matrices with appropriate dimensions. The function $\phi : \mathbb{R}^d \to \mathbb{R}^d$ defines the feedback rule that relates the output of this system to its input.

Consider the GD method for solving problem (2.1). Given $x_0 \in \mathbb{R}^d$, the GD iterations with a constant stepsize $\alpha > 0$ take the form for $k \geq 0$

$$x_{k+1} = x_k - \alpha \nabla f(x_k), \tag{2.3}$$

which can be cast as (2.2) by setting $\xi_k = x_k$, $\phi(\cdot) = \nabla f(\cdot)$ and letting

$$A = I_d, \quad B = -\alpha I_d, \quad C = I_d, \quad D = 0_d. \tag{2.4}$$

On the other hand, when implemented on (2.1), the AG method with constant stepsize $\alpha > 0$ and momentum parameter $\beta > 0$ generates the iterates as follows for $k \geq 0$:

$$y_k = (1 + \beta)x_k - \beta x_{k-1}, \quad x_{k+1} = y_k - \alpha \nabla f(y_k). \tag{2.5}$$

Setting $\phi(\cdot) = \nabla f(\cdot)$ and defining the state vector $\xi_k = \left[\begin{smallmatrix} x_k^\top & x_{k-1}^\top \end{smallmatrix}\right]^\top$, AG iterations can be rewritten as in (2.2) for

$$A = \begin{bmatrix} (1 + \beta) I_d & -\beta I_d \\ I_d & 0_d \end{bmatrix}, \quad B = \begin{bmatrix} -\alpha I_d \\ 0_d \end{bmatrix}, \quad C = \begin{bmatrix} (1 + \beta) I_d & -\beta I_d \end{bmatrix}, \quad D = 0_d. \tag{2.6}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 05/03/20 to 18.9.61.111. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

For both algorithms, the iterates x_k are captured by the state ξ_k of the dynamical system representation.

In this work, we assume that at each iteration $k \geq 0$, instead of the actual gradient $\nabla f(y_k)$, we have access to a noisy version $\nabla f(y_k) + w_k$, where $w_k \in \mathbb{R}^d$ represents the additive noise. In the dynamical system representation, the noisy iterations of the GD and AG algorithms could be written as

$$\xi_{k+1} = A\xi_k + B(u_k + w_k), \quad y_k = C\xi_k, \quad u_k = \nabla f(y_k), \tag{2.7}$$

where A, B, C, and D are selected according to (2.4) for GD or (2.6) for AG.[1] Except for section 5, where we study deterministic perturbations, we assume throughout this paper that the sequence $\{w_k\}_k$ of random variables satisfies the following assumption.

Assumption 2.1. For any $k \geq 0$, the random variable w_k in (2.7) is zero mean and independent from $\{\xi_i\}_{i=1}^k$ and $\{y_i\}_{i=1}^k$. In addition, there exists a scalar $\sigma > 0$ such that $\mathbb{E}(w_k w_k^\top) = \sigma^2 I_d$ for any $k \geq 0$.

This noise structure arises naturally in stochastic optimization where the full gradient is approximated from finitely many samples (see, e.g., [39]), in regression problems [2, 14, 19], as well as in optimization algorithms where the full gradient is subject to an isotropic noise or uncertainty (see, e.g., [30, 31]). The special case when w_k is Gaussian also appears in algorithms where random noise is intentionally injected to the gradient to guarantee privacy (e.g., [3]) or to ensure global convergence, as in the Euler–Mariyama discretization of the overdamped and underdamped Langevin dynamics [8, 16, 21, 22]. It will be clear from our discussion that our results can be extended to the *structured noise* case, i.e., when the covariance matrix $\mathbb{E}(w_k w_k^\top) = S$ for some positive definite matrix S.

Consider a first-order algorithm (e.g., GD or AG) subject to additive noise satisfying Assumption 2.1. For this scenario, where the noise is *persistent*, i.e., it does not decay over time, it is possible that $\lim_{k\to\infty} \mathbb{E}[f(x_k)]$ may *not* exist; therefore, one natural way of defining *robustness* of an algorithm to noise is to consider the worst case limiting suboptimality along all possible subsequences, i.e.,

$$\mathcal{J} \triangleq \limsup_{k\to\infty} \frac{1}{\sigma^2} \mathbb{E}[f(x_k) - f^*]. \tag{2.8}$$

Clearly, $\mathcal{J}$ depends on the choice of algorithm parameters. Moreover, since the limit $\lim_{k\to\infty} f(x_k)$ may not exist when the gradients are perturbed by persistent additive noise, both notions of "convergence" and "convergence rate" are vague. To make these terms more precise in our context, consider the line segment $[f^*,\ f^* + \sigma^2\mathcal{J}]$. In the subsequent sections of the paper, we show that $\{f(x_k)\}_{k\geq 0}$ sequence converges to this line segment *linearly* with a rate depending on the algorithm parameters. Thus, the aim of this paper is to investigate this trade-off between the robustness and rate associated with a given first-order algorithm and to understand the dependence of these key notions of convergence on the choice of algorithm parameters. We believe that achieving this goal would provide important leverage to decision makers to set the parameters in such a way that fits the purpose of the application. We focus on the expected suboptimality $\{\mathbb{E}[f(x_k) - f^*]\}_k$ in the text since this is typically the object of study in the literature for quantifying the performance of similar algorithms.

[1] Although our focus in this paper will be primarily on GD and AG dynamics under noise, it will be clear from our discussion that our ideas naturally extend to many other algorithms that admit such a dynamical system representation, including the heavy-ball and the robust momentum methods [9, 28, 33].

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 05/03/20 to 18.9.61.111. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

It is worth emphasizing that robustness can also be studied in the solution space. Indeed, let $\{x_k\}_{k\geq 0}$ be a random iterate sequence corresponding to (2.7) where $\{w_k\}_k$ models the additive noise sequence and satisfies Assumption 2.1. Due to the noise injected at each step, the sequence $\{x_k\}$ will oscillate around the optimal solution with a nonzero variance; therefore, another natural metric to measure robustness is the worst-case limiting distance to the optimal solution x^* along all possible iterate subsequences, i.e.,

$$\mathcal{J}' \triangleq \limsup_{k\to\infty} \frac{1}{\sigma^2}\mathbb{E}\left[\|x_k - x^*\|^2\right]. \tag{2.9}$$

Similarly, the convergence rate could be defined to be the linear rate that $\{x_k\}_{k\geq 0}$ converges to the ball $\{x \in \mathbb{R}^d : \|x - x^*\|^2 \leq \sigma^2\mathcal{J}'\}$. The quantity $\mathcal{J}'$ can be viewed as the *robustness to noise in terms of iterates* because it is equal to the ratio of the power of the iterates to the power of the input noise, measuring how much a system amplifies input noise. In particular, the smaller this measure is, the more robust a system is under additive random noise.[2] In section 3, we remark that $\mathcal{J}'$ is indeed the H_2 norm of the dynamical system in (2.7) with $C = I$, a notion applicable to both linear and nonlinear systems [20, 44]. Later in section 5, we will use $\mathcal{J}'$ to make some connections between the robustness of a first-order method with its behavior when perturbed from the optimal solution.

3. Quadratic functions. In order to understand the effect of noise on the dynamics, we find it insightful to first focus on the case where the objective function is quadratic. Let $f \in S_{\mu,L}(\mathbb{R}^d)$ be a quadratic function given by $f(x) = \frac{1}{2}x^\top Qx - p^\top x + r$, where Q is symmetric and positive definite with eigenvalues $\{\lambda_i\}_{i=1}^d$ listed in increasing order satisfying $0 < \mu = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_d = L$. The gradient of f is given by

$$\nabla f(x) = Qx - p = Q(x - x^*), \tag{3.1}$$

where $x^* = Q^{-1}p$ is the optimal solution to problem (2.1). Plugging the formula for the gradient $\nabla f(y_k)$ from (3.1) into (2.7), we obtain

$$\xi_{k+1} = (A + BQC)\xi_k - BQx^* + Bw_k, \quad y_k = C\xi_k. \tag{3.2}$$

With ξ^* equal to x^* for GD and $[x^{*\top}\, x^{*\top}]^\top$ for AG, in both cases we have $\xi^* = A\xi^*$ and $x^* = C\xi^*$, where A and C are given in (2.4) and (2.6) for GD and AG, respectively. Therefore, defining $\tilde{y}_k \triangleq y_k - x^*$ and $\tilde{\xi}_k \triangleq \xi_k - \xi^*$, (3.2) yields

$$\tilde{\xi}_{k+1} = A_Q\tilde{\xi}_k + Bw_k, \quad \tilde{y}_k = C\tilde{\xi}_k, \tag{3.3}$$

where A_Q is the state-transition matrix given by $A_Q = A + BQC$.

In the absence of noise (when $w_k = 0$ for all k), if $\rho(A_Q)$ is less than one, then we clearly have $\tilde{\xi}_k \to 0$ and $\tilde{y}_k \to 0$ *linearly*. As a consequence, the suboptimality, $f(x_k) - f^*$, goes to zero linearly as well. On the other hand, when the gradients are perturbed by random additive noise, as we shall discuss in the next section, $\mathbb{E}[f(x_k) - f^*]$ does *not* go to zero.

[2]See Appendix E, provided as a supplementary material, where we derive robustness results based on $\mathcal{J}'$ for both GD and AG.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 05/03/20 to 18.9.61.111. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

3.1. Performance metrics under gradient noise: Rate and robustness. In this section, we use the dynamical system representation of the algorithms given in (3.3) to study the limiting behavior of the expected suboptimality $\mathbb{E}[f(x_k) - f^*]$. We show that this sequence converges, i.e., the limit of the expected suboptimality exists and it is equal to the limit superior in (2.8). We also provide the associated convergence rate and present an explicit characterization of the limiting value using insights from robust control theory. More specifically, consider the shifted state sequence $\{\tilde{\xi}_k\}_k$ generated according to (3.3). Since w_k is zero mean for all $k \geq 0$ (Assumption 2.1), by taking the expectation of (3.3), we obtain

$$\mathbb{E}[\tilde{\xi}_k] = A_Q^k \ \tilde{\xi}_0 \quad \forall\ k \geq 0. \tag{3.4}$$

Therefore, under the assumption that $\rho(A_Q) < 1$, the sequence $\{\mathbb{E}[\tilde{\xi}_k]\}_k$ converges to zero with an asymptotic linear rate $\rho(A_Q)$. Note that the state sequence $\{\xi_k\}_k$ and the iterates $\{x_k\}_k$ can be related by defining $T = I_d$ for GD and $T = [I_d \quad 0_d]$ for AG so that $x_k = T\xi_k$ for all $k \geq 0$.

Recall the robustness definition given in (2.8). In the next lemma, we focus on the suboptimality sequence, $\{f(x_k) - f^*\}_k$ for quadratic f and we show that the limit,

$$\mathcal{J} = \frac{1}{\sigma^2} \lim_{k\to\infty} \mathbb{E}[f(x_k) - f^*], \tag{3.5}$$

exists; moreover, for some $\{\varepsilon_k\}_k \subset [0, \infty)$ such that $\lim_{k\to\infty} \varepsilon_k = 0$, we have

$$\left|\mathbb{E}[f(x_k) - f^*] - \sigma^2 \mathcal{J}\right| \leq \psi_0 \ \left(\rho(A_Q) + \varepsilon_k\right)^{2k} \quad \forall\ k \geq 0, \tag{3.6}$$

where $\rho(A_Q)$ is the spectral radius of A_Q, and ψ_0 is a constant that may depend on the initialization x_0. This shows that the sequence $\{\mathbb{E}[f(x_k) - f^*]\}_k$ converges to an interval around the origin with radius $\sigma^2\mathcal{J}$, and the convergence is linear with an asymptotical rate that is arbitrarily close to $\rho(A_Q)^2$. It is therefore natural to define the normalized radius, $\mathcal{J}$, as *robustness* of the system to gradient noise, i.e., if this radius is bigger, it means that the asymptotic error of the algorithm in terms of the function value is larger; hence, the algorithm is less robust to the injected noise.

The limit in (3.5) can be evaluated by using the tools from *standard H_2 theory* arising in robust control of dynamical systems (see, e.g., [26]) as we shall explain below. The H_2 norm is a well-known fundamental metric for quantifying the robustness of a linear dynamical system to noise in control engineering and has been widely used in designing the parameters of control systems subject to noise. Given arbitrary matrices (A, B, C) and $D = 0_d$, consider a linear system as in (2.2) but *without* feedback ϕ. Suppose there exists ξ^* and y^* such that $\xi^* = A\xi^*$ and $y^* = C\xi^*$. The H_2 norm of this linear system, denoted by $H_2(A, B, C)$, measures the stationary variance of the output response $\{y_k\}$ to unit white noise input [46], i.e.,

$$H_2^2(A, B, C) \triangleq \lim_{k\to\infty} \frac{1}{\sigma^2} \mathbb{E}\|y_k - y^*\|^2. \tag{3.7}$$

The H_2 norm admits alternative definitions, which are all equivalent for linear systems (see, e.g., [44, 46]). When it is clear from the context, we will remove the dependency of the H_2 norm to the system matrices (A, B, C). The H_2 norm can be computed as

$$H_2^2(A, B, C) = \mathrm{Tr}\left(CX_0C^\top\right), \tag{3.8}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 05/03/20 to 18.9.61.111. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

where X_0 solves the *discrete Lyapunov* equation:[3]

$$AX_0A^\top - X_0 + BB^\top = 0 \tag{3.9}$$

(see, e.g., [26, 46]). Moreover, if $BB^\top$ is positive definite and A is *discrete-time stable* (i.e., $\rho(A) < 1$), the solution admits the following formula:

$$X_0 = \sum_{k=0}^{\infty} (A^\top)^k B^\top B A^k \tag{3.10}$$

(see, e.g., [46]). We will show in the following lemma that the limit $\mathcal{J}$ in (3.5) exists for quadratic objectives. Our proof technique is based on relating $\mathcal{J}$ to the H_2 norm of a transformed linear system as follows. We first rewrite the suboptimality $f(x_k) - f^*$ in terms of the iterates x_k:

$$f(x_k) - f^* = \tfrac{1}{2}(x_k - x^*)^\top Q(x_k - x^*) = \tfrac{1}{2}(T\tilde{\xi}_k)^\top Q(T\tilde{\xi}_k) = (RT\tilde{\xi}_k)^\top (RT\tilde{\xi}_k), \tag{3.11}$$

where we used the fact that $x_k = T\xi_k$, and $\frac{1}{2}Q = R^\top R$ is the Cholesky decomposition of $\frac{1}{2}Q$. If we consider the system defined by matrices (A_Q, B, RT), it follows from the definition of the H_2 norm (3.7) and (3.11) that

$$\mathcal{J} = H_2^2(A_Q, B, \tilde{R}), \quad \text{where} \quad \tilde{R} \triangleq RT. \tag{3.12}$$

By (3.8) and (3.9), we have $\mathcal{J} = \mathrm{Tr}(\tilde{R}X\tilde{R}^\top)$, where X is the solution to

$$A_Q X A_Q^\top - X + BB^\top = 0. \tag{3.13}$$

LEMMA 3.1. *Consider the linear dynamical system* (3.3) *defined by the matrices* (A_Q, B, C). *If* $\rho(A_Q) < 1$, *then the limit in* (3.5) *exists, i.e.,* $\sigma^2\mathcal{J} = \lim_{k\to\infty} \mathbb{E}[f(x_k) - f^*]$, *and there exists a nonnegative sequence* $\{\varepsilon_k\}_k$ *such that* $\lim_k \varepsilon_k = 0$ *and*

$$\left|\mathbb{E}[f(x_k) - f^*] - \sigma^2\mathcal{J}\right| \le \ \psi_0 \ (\rho(A_Q) + \varepsilon_k)^{2k} \quad \forall\ k \ge 0$$

holds for some explicitly given positive constant ψ_0 *that depends on the initialization* x_0. *Furthermore, when* A_Q *is symmetric,* $\varepsilon_k = 0$ *for every* $k \ge 0$.

Proof. Using (3.11) and $\mathcal{J} = \mathrm{Tr}(\tilde{R}X\tilde{R}^\top)$, we obtain

$$\begin{aligned} \mathbb{E}[f(x_k) - f^*] - \sigma^2 H_2^2(A_Q, B, \tilde{R}) &= \mathbb{E}[(\tilde{R}\tilde{\xi}_k)^\top(\tilde{R}\tilde{\xi}_k)] - \sigma^2\,\mathrm{Tr}(\tilde{R}X\tilde{R}^\top) \\ &= \mathrm{Tr}(\tilde{R}\mathbb{E}[\tilde{\xi}_k\tilde{\xi}_k^\top]\tilde{R}^\top) - \sigma^2\,\mathrm{Tr}(\tilde{R}X\tilde{R}^\top) \\ &= \mathrm{Tr}(\tilde{R}(V_k - \sigma^2 X)\tilde{R}^\top), \end{aligned} \tag{3.14}$$

where $V_k \triangleq \mathbb{E}[\xi_k\xi_k^\top]$ for $k \ge 0$. It follows from (3.3) that

$$\begin{aligned} V_k = \mathbb{E}[\tilde{\xi}_k\tilde{\xi}_k^\top] &= \mathbb{E}[(A_Q\tilde{\xi}_{k-1} + Bw_{k-1})(A_Q\tilde{\xi}_{k-1} + Bw_{k-1})^\top] \\ &= A_Q V_{k-1} A_Q^\top + \sigma^2 BB^\top \end{aligned} \tag{3.15}$$

holds for all $k \ge 1$, where in the last equality we used the fact that the random vector w_{k-1} is zero mean, independent of ξ_{k-1}, and has covariance matrix $\mathbb{E}[w_{k-1}w_{k-1}^T] = \sigma^2 I_d$. Moreover, by (3.13) we have $X = A_Q X A_Q^T + BB^T$; hence, subtracting $\sigma^2 X$ from both sides of (3.15), we obtain

[3] The value of H_2^2 can also be computed as $\mathrm{Tr}(B^\top\tilde{X}_0 B)$, where $\tilde{X}_0$ solves $A^\top\tilde{X}_0 A - \tilde{X}_0 + C^\top C = 0$.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

$$V_k - \sigma^2 X = A_Q(V_{k-1} - \sigma^2 X)A_Q^\top = A_Q^k(V_0 - \sigma^2 X)(A_Q^\top)^k, \tag{3.16}$$

where the last equality comes from recursively using the first equality. This implies

$$\begin{aligned} |\operatorname{Tr}(\tilde{R}(V_k - \sigma^2 X)\tilde{R}^\top)| &= |\operatorname{Tr}(\tilde{R}A_Q^k(V_0 - \sigma^2 X)(A_Q^\top)^k\tilde{R}^\top)| \\ &= |\operatorname{Tr}((V_0 - \sigma^2 X)(\tilde{R}A_Q^k)^\top(\tilde{R}A_Q^k))| \\ &\le m\|V_0 - \sigma^2 X\|\|\tilde{R}A_Q^k\|^2 \le m\|V_0 - \sigma^2 X\|\|\tilde{R}\|^2\|A_Q^k\|^2, \end{aligned} \tag{3.17}$$

where $\|.\|$ is the spectral norm, and the first inequality in (3.17) follows from the Von Neumann's trace inequality, which states that for any two $m \times m$ matrices U and V with singular values $||U||_2 = u_1 \ge \cdots \ge u_m$ and $||V||_2 = v_1 \ge \cdots \ge v_m$, respectively, we have $|\operatorname{Tr}(UV)| \le \sum_{i=1}^m u_i v_i$. Finally, it follows from the Gelfand's formula that there exists a sequence of nonnegative numbers $\{\varepsilon_k\}_k$ such that for every $k \ge 0$, $\|A_Q^k\|_2 \le (\rho(A_Q) + \varepsilon_k)^k$ and $\lim_k \varepsilon_k = 0$. Note that when A_Q is symmetric, we have $\|A_Q^k\|_2 = \rho(A_Q)^k$ so that we can choose $\varepsilon_k = 0$. Inserting this bound into (3.17), we obtain the desired result. ∎

It is worth noting that for a strongly convex quadratic function in the form of $f(x) = \frac{1}{2}x^\top Qx - p^\top x + r$, a similar line of argument as in Lemma 3.1 shows that $\mathcal{J}'$ in (2.9) is in fact equal to $\lim_{k\to\infty} \frac{1}{\sigma^2}\mathbb{E}[\|x_k - x^*\|^2] = H_2^2(A_Q, B, T)$. Next we focus on the GD and AG algorithms, discuss the dependence of their convergence rate and robustness on the parameters (stepsize α and momentum β), and show how to formulate an optimization problem that systematically trades off convergence rate and robustness.

3.2. Gradient descent method. The dynamical system representation of GD, choosing the A, B, C as in (2.4), yields

$$\mathbb{E}[\xi_{k+1}] = A_Q\mathbb{E}[\xi_k] \quad \text{with} \quad A_Q = I_d - \alpha Q.$$

As shown in Lemma 3.1, the convergence rate of GD is given by $\rho(A_Q)^2$. For GD, we will suppress the dependence of $\rho(A_Q)$ on A_Q and use the notation $\rho(\alpha)$ to highlight the effect of the stepsize α. Since A_Q is symmetric, $\rho(\alpha)$ can be computed as

$$\rho(\alpha) = \rho(A_Q) = \|A_Q\| = \max\{|1 - \alpha\mu|, |1 - \alpha L|\}. \tag{3.18}$$

$\alpha \in (0, 2/L)$ is a necessary condition for global linear convergence; otherwise, $\rho(\alpha) \ge 1$. In particular, it is well-known that the fastest rate is achieved for the stepsize

$$\bar{\alpha} \triangleq \arg\min_{\alpha\ge 0} \rho(A_Q) = \frac{2}{\mu + L}, \tag{3.19}$$

which leads to a convergence rate of $\bar{\rho} = 1 - \frac{2}{\kappa+1}$. The choice of the stepsize affects not only the rate (see (3.18)) but also the robustness of the GD algorithm to gradient noise. The following proposition provides an analytical characterization of the robustness $\mathcal{J}$ of the GD method as a function of the stepsize, which we denote by $\mathcal{J}(\alpha)$ to highlight its dependence on α.

PROPOSITION 3.2. *Let f be a quadratic function of the form $f(x) = \frac{1}{2}x^\top Qx - p^\top x + r$. Consider the GD iterations given by (2.3) with constant stepsize $\alpha \in (0, 2/L)$. Then the robustness of the GD method is given by*

$$\mathcal{J}(\alpha) = \sum_{i=1}^d \frac{\alpha^2\lambda_i}{2(1 - (1 - \alpha\lambda_i)^2)} = \alpha\sum_{i=1}^d \frac{1}{2(2 - \alpha\lambda_i)}, \tag{3.20}$$

where $0 < \mu = \lambda_1 \le \lambda_2 \le \cdots \lambda_d = L$ are the eigenvalues of Q.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 05/03/20 to 18.9.61.111. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Downloaded 05/03/20 to 18.9.61.111. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Proof. We first show that without loss of generality we can assume Q is a diagonal matrix. Let $Q = U\Lambda U^\top$ be the eigenvalue decomposition of Q where U is a unitary matrix and $\Lambda = \mathbf{diag}(\lambda_1, \dots, \lambda_d)$ is a diagonal matrix containing the eigenvalues of Q. Multiplying A_Q by $U^\top$ and U from left and right leads to

$$U^\top A_Q U = U^\top (I_d - \alpha Q)U = I_d - \alpha\Lambda = A_\Lambda, \tag{3.21}$$

where $A_\Lambda \triangleq I_d - \alpha\Lambda$ is a diagonal matrix. Similarly, we multiply the Lyapunov equation (3.13) from left and right by $U^\top$ and U, which yields $U^\top A_Q X A_Q^\top U - U^\top X U + \alpha^2 I_d = 0$, where we have used the fact that $B = -\alpha I_d$ for the dynamical system representation of the GD method (see (2.4)). It follows from (3.21) that $A_Q = U(I_d - \alpha\Lambda)U^\top$, which when plugged into the Lyapunov equation above yields $(I_d - \alpha\Lambda)U^\top X U(I_d - \alpha\Lambda) - U^\top X U + \alpha^2 I_d = 0$. This means that the matrix $U^\top X U$ solves the Lyapunov equation obtained by replacing A_Q by A_Λ in (3.13). Furthermore, the Cholesky decomposition of $\frac{1}{2}\Lambda$ is equal to $\sqrt{\frac{1}{2}}\Lambda^{1/2}$; thus, the robustness $\mathcal{J}$, corresponding to $H_2^2(A_\Lambda, B, \sqrt{\frac{1}{2}}\Lambda^{1/2}T)$, is equal to

$$\frac{1}{2}\operatorname{Tr}(\Lambda^{1/2}TU^\top X U(\Lambda^{1/2}T)^\top) = \frac{1}{2}\operatorname{Tr}(\Lambda^{1/2}U^\top X U\Lambda^{1/2}) = \operatorname{Tr}(\tilde{R}X\tilde{R}^\top),$$

where we used $T = I_d$ for GD for the first equality and the fact that the Cholesky decomposition of $\frac{1}{2}Q$ is $(\sqrt{\frac{1}{2}}\Lambda^{1/2}U^\top)^\top(\sqrt{\frac{1}{2}}\Lambda^{1/2}U^\top)$ to obtain the second equality. Therefore, robustness $\mathcal{J}$ would be invariant if we were to replace Q by Λ and solve the Lyapunov equation (3.13) for (A_Λ, B) instead of (A_Q, B). With this replacement, it is easy to verify that the solution of the Lyapunov equation is $X_\Lambda = \mathbf{diag}(\frac{\alpha^2}{1-(1-\alpha\lambda_1)^2}, \dots, \frac{\alpha^2}{1-(1-\alpha\lambda_d)^2}))$ as A_Λ and B are both diagonal. Plugging this solution into $\frac{1}{2}\operatorname{Tr}(\Lambda^{1/2}X_\Lambda\Lambda^{1/2})$ implies $\mathcal{J}(\alpha) = \sum_{i=1}^d \frac{\alpha^2\lambda_i}{2(1-(1-\alpha\lambda_i)^2)} = \alpha\sum_{i=1}^d \frac{1}{2(2-\alpha\lambda_i)}$, which completes the proof. ∎

Remark 3.1. Proposition 3.2 also shows that the robustness $\mathcal{J}(\alpha)$ for the GD method is an increasing function of α. This means choosing a smaller stepsize leads to GD being more robust, which has been previously observed in the literature for both additive and multiplicative deterministic noise [18, 33].

Having explicit expressions for both convergence rate and robustness for GD (see (3.18) and (3.20)), given an allowable deviation $\epsilon > 0$ from the optimal convergence rate $\bar{\rho} = 1 - \frac{2}{\kappa+1}$, a natural approach to account for the trade-off between these two measures is to choose the stepsize α that results in the most robust algorithm satisfying the rate constraints, i.e., optimizing

$$\min_{\alpha\in(0,2/L)} \mathcal{J}(\alpha) \quad \text{subject to} \quad \rho(\alpha) \le (1+\epsilon)\bar{\rho}. \tag{3.22}$$

This problem is equivalent to the following convex problem for $\epsilon \in [0, \frac{2}{\kappa-1})$ (which ensures that the upper bound on the rate is less than one and the optimization problem (3.22) admits a solution):

$$\min_{\alpha\in(0,2/L)} \mathcal{J}(\alpha) \quad \text{subject to} \quad \frac{1}{1-\rho^2(\alpha)} \le \frac{1}{1-(1+\epsilon)^2\bar{\rho}^2}. \tag{3.23}$$

Indeed, $1/(1-\rho^2)$ is a nondecreasing convex function for $\rho \in (0,1)$ and $\rho(\alpha)$ is convex in α; therefore, both $1/(1-\rho(\alpha)^2)$ and $\mathcal{J}(\alpha)$ in (3.20) are convex for $\alpha \in (0, \frac{2}{L})$ and are

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 05/03/20 to 18.9.61.111. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

increasing in α. Moreover, (3.23) satisfies the Slater condition. Thus, strong duality implies that there exists τ (which is a function of ϵ) such that the above minimization problem is equivalent to the following unconstrained problem:

$$\alpha_*(\tau) \triangleq \underset{\alpha\in(0,2/L)}{\arg\min}\ F_\tau(\alpha) \triangleq \mathcal{J}(\alpha) + \tau\frac{1}{1-\rho^2(\alpha)}. \tag{3.24}$$

The parameter $\tau > 0$ determines the trade-off between rate and robustness. For small τ, the dominant term in the cost would be $\mathcal{J}(\alpha)$ so that we expect the optimal stepsize to be small since $\mathcal{J}(\alpha)$ is an increasing function of α. On the other hand, for large enough τ, the convergence rate is the dominant term in the cost; therefore, one would expect the optimal stepsize (which solves the problem (3.24)) to be close to $\bar{\alpha}$ which corresponds to the fastest achievable rate $\bar{\rho}$ (see (3.19)). In order to get more intuition about the effect of the choice of the stepsize parameter, we next give an illustrative example in dimension $d = 2$ to show the behavior of the optimal $\alpha_*(\tau)$ as the trade-off parameter τ is varied from zero to infinity. For computational tractability, we consider the unconstrained version of the problem given in (3.24).[4]

Example 3.2. In dimension $d = 2$, let $\tau = 2$ and consider the parameters

$$\mu = \lambda_1 = 0.1 \quad \text{and} \quad L = \lambda_2 = 1 \quad \text{with} \quad \kappa = \frac{L}{\mu} = 10. \tag{3.25}$$

The first-order optimality conditions for (3.24) are derived in Proposition A.1, which is equivalent to a polynomial root finding problem in α for a polynomial of degree 4. The roots of polynomials can be found up to arbitrary accuracy by calculating the eigenvalues of the corresponding companion matrix [17], for instance, using the `roots` function in MATLAB. After a careful examination of all the roots, we conclude that the optimal stepsize α_* that minimizes the cost $F_\tau(\alpha)$ is $\alpha_* \approx 1.5055$, which gives the rate $\rho(\alpha_*) \approx 0.8494$ and robustness $\mathcal{J}(\alpha_*) \approx 1.9294$. This point is marked on Figure 1, which shows the robustness level as a function of the optimal convergence rate ρ when we change τ from zero (corresponds to the rightmost point in the curve) to infinity (corresponds to the uppermost point in the curve) for the parameters in (3.25).

The left and middle panels of Figure 1 show the convergence rate and robustness corresponding to the optimal stepsize α_* as a function of the trade-off parameter τ. As τ goes to 0, the robustness term is more dominant, which requires a smaller stepsize;

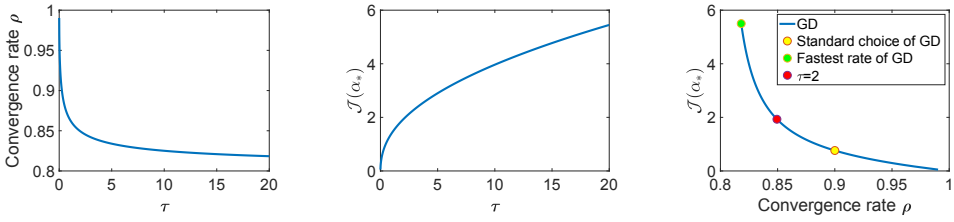

FIG. 1. *Left and middle: The behaviors of the convergence rate ρ and robustness $\mathcal{J}$ computed at the optimal stepsize α_*, as a function of the trade-off parameter τ. Right: The robustness level as a function of convergence rate again as τ varies from zero to infinity.*

[4]In Proposition A.1 of the appendix, we derive the first-order conditions for $\alpha_*(\tau)$ that allow it to be computed up to an arbitrary accuracy.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 05/03/20 to 18.9.61.111. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

therefore, α_* goes to 0 and $\rho(\alpha_*)$ thus goes to 1. As τ becomes larger, the convergence rate becomes more important, and the stepsize also becomes larger to ensure faster convergence. In particular, as τ goes to infinity, α_* goes to $\bar{\alpha}$ given in (3.19), leading to the fastest rate, $\bar{\rho} = \frac{\kappa-1}{\kappa+1} \approx 0.8182$.

Finally, the rightmost panel of Figure 1 illustrates the trade-off between the rate and robustness. We see that for small τ, the optimal stepsize α_* is smaller, which implies improved robustness but slower convergence. As τ grows, we achieve a faster rate at the expense of being less robust to the additive gradient noise. In addition, the points corresponding to the fastest rate, i.e., $\alpha = 2/(\mu + L)$, and standard parameter choice $\alpha = 1/L$ for GD have been marked on this trade-off curve.

We see from Figure 1 that smaller values of ρ (or equivalently smaller values of $\frac{1}{1-\rho^2}$) are accompanied by larger values of $\mathcal{J}$. This suggests that the product $\mathcal{J}\frac{1}{1-\rho^2}$ cannot be too small for any choice of the stepsize α. The next lemma shows that there are some fundamental limits (lower bounds) on how robust the GD can be.

PROPOSITION 3.3. *Let $\rho(\alpha)$ and $\mathcal{J}(\alpha)$ be given by* (3.18) *and* (3.20)*, respectively. Then, the inequality $\mathcal{J}(\alpha) \geq (1-\rho^2(\alpha)) \sum_{i=1}^{d} \frac{1}{8\lambda_i}$ holds for any choice of the stepsize $\alpha > 0$.*

Proof. It follows from (3.18) that for every $i \in \{1, \dots, d\}$, we have $\rho(\alpha) \geq |1-\alpha\lambda_i|$. This implies that $\frac{1}{1-\rho(\alpha)^2} \geq \frac{1}{1-(1-\alpha\lambda_i)^2}$. Multiplying both sides by $\frac{\alpha^2\lambda_i}{2(1-(1-\alpha\lambda_i)^2)}$ and summing over all i yields $\frac{1}{1-\rho^2}\sum_{i=1}^{d} \frac{\alpha^2\lambda_i}{2(1-(1-\alpha\lambda_i)^2)} \geq \sum_{i=1}^{d} \frac{\alpha^2\lambda_i}{2(1-(1-\alpha\lambda_i)^2)^2}$. Given the explicit characterization of $\mathcal{J}(\alpha)$ in Proposition 3.2 (see (3.20)) we obtain

$$\frac{1}{1-\rho^2}\mathcal{J}(\alpha) \geq \frac{1}{2}\sum_{i=1}^{d} \frac{\alpha^2\lambda_i}{(1-(1-\alpha\lambda_i)^2)^2}. \tag{3.26}$$

The right-hand side of (3.26) admits a lower bound as follows:

$$\frac{1}{2}\sum_{i=1}^{d} \frac{\alpha^2\lambda_i}{(1-(1-\alpha\lambda_i)^2)^2} = \frac{1}{2}\sum_{i=1}^{d} \frac{\alpha^2\lambda_i}{(\alpha\lambda_i(2-\alpha\lambda_i))^2} = \frac{1}{2}\sum_{i=1}^{d} \frac{1}{\lambda_i(2-\alpha\lambda_i)^2} \geq \sum_{i=1}^{d} \frac{1}{8\lambda_i}, \tag{3.27}$$

where the last inequality follows from the fact that $|2-\alpha\lambda_i| \leq 2$. Using the lower bound (3.27) along with (3.26) completes the proof. ∎

3.3. Accelerated gradient method. The dynamical system representation of AG, given A, B, C in (2.6), leads to

$$A_Q = \begin{bmatrix} (1+\beta)(I_d - \alpha Q) & -\beta(I_d - \alpha Q) \\ I_d & 0_d \end{bmatrix}. \tag{3.28}$$

We will first formulate an analogous problem to (3.24) for the AG method to design the parameters (α, β) in a way to find a trade-off between the rate and the robustness. Because AG has the pair (α, β) as design parameters, the analogue of (3.24) is

$$(\alpha_*, \beta_*) \triangleq \arg\min_{(\alpha,\beta)\in\mathcal{S}} F_\tau(\alpha, \beta) \triangleq \mathcal{J}(\alpha, \beta) + \tau\frac{1}{1-\rho(\alpha,\beta)^2}, \tag{3.29}$$

where $\mathcal{J}(\alpha, \beta)$ is the robustness to the noise for the system (3.3), $\rho(\alpha, \beta)$ is the convergence rate of AG with parameters (α, β), and $\mathcal{S}$ is the set of all possible choices of the tuple (α, β) so that the AG iterations are globally convergent, i.e.,

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

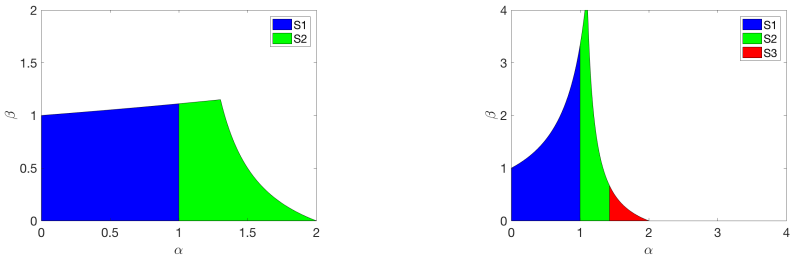

FIG. 2. *Left: The stability region* $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2 \cup \mathcal{S}_3$ *with parameters* $\mu = 0.7$ *and* $L = 1$. *Right: The stability region* $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2$ *with parameters* $\mu = 0.1$ *and* $L = 1$.

$$\mathcal{S} = \{(\alpha, \beta) \ : \ \rho(A_Q) < 1, \alpha \geq 0, \beta \geq 0\} \subset \mathbb{R}^2. \tag{3.30}$$

We call $\mathcal{S}$ the *stability region* of AG, in analogy with the stability region of numerical methods that arise in the discretization of continuous-time differential equations.

We next provide an explicit characterization for the convergence rate and robustness of AG for any given parameters $(\alpha, \beta) \in \mathcal{S}$. The convergence rate ρ of the AG method as a function of α and β is well-known. Diagonalizing the A_Q matrix using the eigenvalue decomposition of Q, it can be shown after some computations that the rate $\rho = \rho(\alpha, \beta)$ admits the formula

$$\rho(\alpha, \beta) = \rho(A_Q) = \max\{\rho_\mu(\alpha, \beta), \rho_L(\alpha, \beta)\}, \tag{3.31}$$

where A_Q is defined by (3.28) and ρ_λ is defined for $\lambda \in \{\mu, L\}$ as follows:

$$\rho_\lambda(\alpha, \beta) = \begin{cases} \frac{1}{2}|(1+\beta)(1-\alpha\lambda)| + \frac{1}{2}\sqrt{\Delta_\lambda} & \text{if } \Delta_\lambda \geq 0, \\ \sqrt{\beta(1-\alpha\lambda)} & \text{otherwise,} \end{cases} \tag{3.32}$$

$$\Delta_\lambda = (1+\beta)^2(1-\alpha\lambda)^2 - 4\beta(1-\alpha\lambda)$$

(see, e.g., [33, Appendix A], [38, section 4.3]). The explicit expression (3.31) for the rate allows us to characterize the set $\mathcal{S}$ in the next proposition, whose proof can be found in the appendix. We illustrate the set $\mathcal{S}$ in Figure 2 for different choices of the parameters μ and L.[5]

PROPOSITION 3.4. *Let* $\mathcal{S}$ *be the stability set of Nesterov's accelerated method defined by* (3.30). *Then its closure is given by the union of the following three sets:*

$$\begin{aligned} \mathcal{S}_1 &:= \left\{(\alpha, \beta) : 0 \leq \alpha \leq \frac{1}{L}, \ 0 \leq \beta(1-\alpha\mu) \leq 1\right\}, \\ \mathcal{S}_2 &:= \left\{(\alpha, \beta) : \frac{1}{L} < \alpha \leq \min\left\{\frac{2}{L}, \frac{1}{\mu}\right\}, \alpha L - 1 \leq \frac{1}{2\beta+1}, \beta(1-\alpha\mu) \leq 1\right\}, \\ \mathcal{S}_3 &:= \left\{(\alpha, \beta) : \frac{1}{\mu} \leq \alpha \leq \frac{2}{L}, \alpha L - 1 \leq \frac{1}{2\beta+1}\right\}, \end{aligned} \tag{3.33}$$

with the convention that $\mathcal{S}_3$ *is the empty set if* $\mu < \frac{L}{2}$.

[5] We note that the stability region of a second-order difference equation that arises in accelerated algorithms that are sublinearly convergent for weakly convex quadratic functions has been studied in [19]; however, these results do not apply to the set $\mathcal{S}$ as we do not require the rate to be accelerated (we consider not only accelerated rates but also any rate ρ less than one) and we consider strongly convex functions instead of weakly convex functions.

Downloaded 05/03/20 to 18.9.61.111. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 05/03/20 to 18.9.61.111. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

The next proposition gives a characterization of the robustness $\mathcal{J}(\alpha,\beta)$ of AG, whose proof can be found in Appendix C.

PROPOSITION 3.5. *Let f be a quadratic function of the form $f(x) = \frac{1}{2}x^\top Qx - p^\top x + r$. Consider the AG iterations given by (2.5) with parameters $(\alpha,\beta) \in \mathcal{S}$. Then the robustness of the AG method is given by*

$$\mathcal{J}(\alpha,\beta) = \sum_{i=1}^{d} u_{\alpha,\beta}(\lambda_i), \tag{3.34}$$

where $\mu = \lambda_1 \le \lambda_2 \le \cdots \le \lambda_d = L$ are the eigenvalues of Q and

$$u_{\alpha,\beta}(\lambda) \triangleq \alpha \frac{1+\beta(1-\alpha\lambda)}{2(1-\beta(1-\alpha\lambda))(2+2\beta-\alpha\lambda(1+2\beta))}. \tag{3.35}$$

In the special case, choosing $\beta = 0$ reduces to the formula (3.20) derived for GD.

Since we have an exact characterization of $\mathcal{J}(\alpha,\beta)$, we can derive the optimality conditions for the problem (3.29) by an approach similar to Proposition A.1, where the optimizer can be characterized as a root of some polynomial. In dimension $d = 2$, given parameters μ and L, the optimizer is easy to compute. However, in high dimensions, this is computationally expensive as it would require determining all the eigenvalues of Q, which can be as expensive as optimizing the objective function f. Nevertheless, exploiting the convexity properties of the function $u_{\alpha,\beta}(\lambda)$, we develop a tractable upper bound for $\mathcal{J}(\alpha,\beta)$ that only depends on μ and L, hence is tractable. Moreover, in the numerical experiments section, we present experiments illustrating that this approach can lead to good performance in terms of trading the speed and the robustness of an algorithm.

To develop this upper bound, first, we show in Lemma D.1 that the function $u_{\alpha,\beta}(\lambda)$ defined in (3.35) is convex in $\lambda \in [\mu, L]$ for fixed $(\alpha,\beta) \in \mathcal{S}$. Therefore, its maximum is attained at one of the endpoints of this interval, i.e.,

$$u_{\alpha,\beta}(\lambda) \le \bar{u}_{\alpha,\beta} \triangleq \max\left[u_{\alpha,\beta}(\mu), u_{\alpha,\beta}(L)\right] \quad \text{for} \quad \lambda \in [\mu, L].$$

Substituting this upper bound in (3.34) and (3.29) leads to

$$\mathcal{J}(\alpha,\beta) \le \bar{\mathcal{J}}(\alpha,\beta) \triangleq d\bar{u}_{\alpha,\beta} \tag{3.36}$$

and the relaxed optimization problem is

$$(\alpha_*,\beta_*) \triangleq \arg\min_{(\alpha,\beta)\in\mathcal{S}} \bar{F}_\tau(\alpha,\beta) \triangleq \bar{\mathcal{J}}(\alpha,\beta) + \tau\frac{1}{1-\rho(\alpha,\beta)^2}. \tag{3.37}$$

This objective only depends on μ and L and is differentiable everywhere in the interior of the stability region $\mathcal{S}$ except when the first term is not differentiable, i.e., when $u_{\alpha,\beta}(\mu) = u_{\alpha,\beta}(L)$, or the second term is not differentiable, i.e., when $\rho_\mu = \rho_L$ or $\Delta_\mu = 0$ or $\Delta_L = 0$. Furthermore, following a similar approach as in Example 3.2, the first-order optimality conditions with respect to α and β result in low-order polynomials (which are independent of the dimension d) which can be solved efficiently up to any accuracy. Thus, other than checking the nondifferentiable points of $\bar{F}_\tau$, the bottleneck in computational complexity is determined by computing the roots of a polynomial with a small degree (whose degree is independent from the dimension d), which is easy to compute even in high dimensions.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 05/03/20 to 18.9.61.111. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

4. Strongly convex functions. In this section, we generalize our analysis to smooth strongly convex functions $f \in S_{\mu,L}(\mathbb{R}^d)$. Note that we can rewrite (2.7) as

$$\xi_{k+1} = A\xi_k + B(\nabla f(y_k) + w_k), \quad y_k = C\xi_k, \tag{4.1}$$

where $w_k \in \mathbb{R}^d$ models the additive noise and satisfies Assumption 2.1. Similar to the previous section, we define $\tilde{y}_k \triangleq y_k - x^*$ and $\tilde{\xi}_k \triangleq \xi_k - \xi^*$, where ξ^* is equal to x^* for GD and $[x^{*\top}\ x^{*\top}]^\top$ for AG, and in both cases we have $\xi^* = A\xi^*$ and $x^* = C\xi^*$, where A and C are given in (2.4) and (2.6) for GD and AG, respectively. We use the equation $x_k = T\xi_k$ to relate x_k and ξ_k for any $k \geq 0$, where $T = I_d$ for GD and $T = [I_d \quad 0_d]$ for AG. To simplify the notation, we define $\bar{f} : \mathbb{R}^m \to \mathbb{R}$ such that $\bar{f}(\xi) = f(T\xi)$ for all $\xi \in \mathbb{R}^m$, which means $\bar{f}(\xi_k) = f(x_k)$ for all $k \geq 0$.

4.1. Rate and robustness. The goal is to extend the definitions of rate and robustness from the quadratic case to general strongly convex functions. We will use

$$\mathcal{J} = \limsup_{k\to\infty} \frac{1}{\sigma^2}\mathbb{E}[f(x_k) - f^*] \tag{4.2}$$

(provided also in (2.8)) to define the robustness of an algorithm and study the convergence rate of the expected suboptimality to an interval around zero with radius $\sigma^2\mathcal{J}$. For both GD and AG, our main results provide upper bounds of the form

$$\mathbb{E}[f(x_k) - f^*] \leq \rho^{2k}\psi_0 + \sigma^2 R \quad \forall\ k \geq 0, \tag{4.3}$$

where ψ_0, R, and $0 < \rho < 1$ are nonnegative numbers all of which depend on algorithm parameters and the initial point x_0. Clearly, R is an upper bound on $\mathcal{J}$; we will show in this section that our bounds are tight. Moreover, we also recover the fastest known rates in the literature in the absence of noise (σ=0). Our upper bounds only depend on μ and L and are computationally tractable and explicit in some cases. With these upper bounds, one can formulate an optimization problem similar to that of the previous section to find the algorithm parameters that can achieve a particular trade-off between rate and robustness.

4.2. Rate and robustness trade-off analysis using Lyapunov functions. We use a *Lyapunov function* approach to provide a bound as in (4.3) for both GD and AG methods. In particular, we consider a family of Lyapunov functions parameterized by a nonnegative constant c and a positive semidefinite matrix P as

$$V_{P,c}(\xi) \triangleq V_P(\xi) + c(\bar{f}(\xi) - f^*), \tag{4.4}$$

where $V_P(\xi) \triangleq (\xi - \xi^*)^\top P(\xi - \xi^*)$, and study the change in the Lyapunov function $V_{P,c}(\xi)$ along $\{\xi_k\}_k$ generated by the dynamical system representation (4.1).

Our first result shows how for some positive semidefinite matrix P and $c = 0$, $V_p(\xi) \triangleq V_{P,0}(\xi)$ evolves along the iterations and provide a characterization of the difference $\mathbb{E}[V_P(\xi_{k+1})] - \rho^2\mathbb{E}[V_P(\xi_k)]$ for any $k \geq 0$ and $\rho \geq 0$.

LEMMA 4.1. *Consider the Lyapunov function $V_P(\xi) = (\xi - \xi^*)^\top P(\xi - \xi^*)$, where $P \succeq 0$. Then, we have*

$$\mathbb{E}[V_P(\xi_{k+1})] = \mathbb{E}\left[\begin{bmatrix}\xi_k - \xi^* \\ \nabla f(y_k)\end{bmatrix}^\top \begin{bmatrix}A^\top PA & A^\top PB \\ B^\top PA & B^\top PB\end{bmatrix}\begin{bmatrix}\xi_k - \xi^* \\ \nabla f(y_k)\end{bmatrix}\right] + \sigma^2 \operatorname{Tr}(B^\top PB). \tag{4.5}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 05/03/20 to 18.9.61.111. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Proof. Since $\xi^* = A\xi^*$, (4.1) implies $\tilde{\xi}_{k+1} = A\tilde{\xi}_k + B(\nabla f(y_k) + w_k)$ for $k \geq 0$. Therefore,

$$\mathbb{E}[V_P(\xi_{k+1})] = \mathbb{E}[\tilde{\xi}_{k+1}^\top P \tilde{\xi}_{k+1}] = \mathbb{E}[(A\tilde{\xi}_k + B(\nabla f(y_k)+w_k))^\top P(A\tilde{\xi}_k + B(\nabla f(y_k)+w_k))]$$

$$= \mathbb{E}\left[\begin{bmatrix}\tilde{\xi}_k \\ \nabla f(y_k)\end{bmatrix}^\top \begin{bmatrix}A^\top PA & A^\top PB \\ B^\top PA & B^\top PB\end{bmatrix}\begin{bmatrix}\tilde{\xi}_k \\ \nabla f(y_k)\end{bmatrix}\right] + \sigma^2 \operatorname{Tr}(B^\top PB). \tag{4.6}$$

where we use the fact that w_k is zero mean and independent from $\tilde{\xi}_k$ and y_k and the assumption that $\mathbb{E}(w_k w_k^T) = \sigma^2 I_d$ to derive (4.6). ∎

COROLLARY 4.2. *For any $\rho \in (0,1)$ and $P \in \mathbb{S}_+^m$, $V_P(\xi)$ satisfies*

$$\mathbb{E}[V_P(\xi_{k+1})] - \rho^2 \mathbb{E}[V_P(\xi_k)] = \mathbb{E}\left[\begin{bmatrix}\xi_k - \xi^* \\ \nabla f(y_k)\end{bmatrix}^\top \Phi(A,B,P,\rho)\begin{bmatrix}\xi_k - \xi^* \\ \nabla f(y_k)\end{bmatrix}\right] + \sigma^2 \operatorname{Tr}(B^\top PB), \tag{4.7}$$

where

$$\Phi(A,B,P,\rho) \triangleq \begin{bmatrix}A^\top PA - \rho^2 P & A^\top PB \\ B^\top PA & B^\top PB\end{bmatrix}. \tag{4.8}$$

We next show how to provide upper bounds on the improvement in $\mathbb{E}[V_{P,c}(\xi_k)]$ by imposing MIs. In particular, given A, B, and C defining the first-order algorithm, we assume there exists a symmetric matrix $X \in \mathbb{S}^{m+d}$ such that

$$X \succeq \Phi(A,B,P,\rho) \tag{4.9}$$

for some $P \in \mathbb{S}_+^m$ and $\rho \in (0,1)$. Moreover, we assume that for some nonnegative constants Γ and c, the same ρ and X satisfy

$$\mathbb{E}\left[\begin{bmatrix}\xi_k - \xi^* \\ \nabla f(y_k)\end{bmatrix}^\top X \begin{bmatrix}\xi_k - \xi^* \\ \nabla f(y_k)\end{bmatrix}\right] \leq c(\rho^2 \mathbb{E}[\bar{f}(\xi_k) - f^*] - \mathbb{E}[\bar{f}(\xi_{k+1}) - f^*] + \sigma^2\Gamma) \tag{4.10}$$

for every $k \geq 0$; hence, it follows from (4.9) and (4.10) along with (4.7) that

$$\rho^2 \mathbb{E}[V_{P,c}(\xi_k)] + \sigma^2(\operatorname{Tr}(B^\top PB) + c\Gamma) \geq \mathbb{E}[V_{P,c}(\xi_{k+1})] \quad \forall\, k \geq 0.$$

Thus, for all $k \geq 0$, we have

$$\mathbb{E}[V_{P,c}(\xi_k)] \leq \rho^{2k} V_{P,c}(\xi_0) + \frac{1-\rho^{2k}}{1-\rho^2}\sigma^2 R_P, \quad R_P \triangleq \operatorname{Tr}(B^\top PB) + c\Gamma. \tag{4.11}$$

This MI-based approach has been used in the literature to study the convergence rate of first-order methods, e.g., [28, 33]. Here we use it to characterize their rate and robustness under additive gradient noise.

4.3. Gradient descent method for strongly convex functions. Recall the GD update rule

$$x_{k+1} = x_k - \alpha(\nabla f(x_k) + w_k), \tag{4.12}$$

which admits the dynamical system representation in (4.1) with A, B, C as in (2.4). The next theorem extends the result of Proposition 3.2 to general strongly convex functions and characterizes the behavior of $\{x_k\}_k$ under additive gradient error.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 05/03/20 to 18.9.61.111. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

PROPOSITION 4.3. *Let $f \in S_{\mu,L}(\mathbb{R}^d)$, and consider the GD iterations given by (4.12). Assume there exist $\rho \in (0,1)$ and $p > 0$ such that*

$$X_0 \succeq \Phi(A, B, P, \rho) \tag{4.13}$$

holds, where $X_0 = \left[\begin{smallmatrix} 2\mu L I_d & -(\mu+L)I_d \\ -(\mu+L)I_d & 2I_d \end{smallmatrix}\right]$ and $P = pI_d$. Then for all $k \geq 0$,

$$\mathbb{E}[\|x_k - x^*\|^2] \leq \rho^{2k} \|x_0 - x^*\|^2 + \frac{1-\rho^{2k}}{1-\rho^2}\, \sigma^2 \alpha^2 d. \tag{4.14}$$

Proof. Noting that $\xi_k = y_k$ for GD, it follows from (1.2) with $x = \xi_k$ and $y = x^*$ that (4.10) holds for $X = X_0$ and $c = 0$. Moreover, (4.13) implies that (4.9) holds for $X = X_0$ and $P = p \otimes I_d$; therefore, (4.11) yields

$$p\mathbb{E}[\|x_k - x^*\|^2] \leq \rho^{2k} p \|x_0 - x^*\|^2 + \frac{1-\rho^{2k}}{1-\rho^2}\sigma^2 \operatorname{Tr}(B^\top P B) \quad \forall\ k \geq 0. \tag{4.15}$$

With $B = -\alpha I_d$ for GD, we have $\operatorname{Tr}(B^\top P B) = \alpha^2 pd$ and this completes the proof. ∎

Note that for a fixed α, a smaller ρ makes both terms of (4.14) smaller as $\frac{1-\rho^{2k}}{1-\rho^2}$ is an increasing function of ρ. If $\alpha \in (0, 2/L)$, it was shown in [33] that there exist (p, ρ) such that the MI in (4.13) holds; moreover, for a given α fixed, the smallest $\rho \in (0,1)$ for which such a positive p exists is equal to

$$\rho_{GD}(\alpha) = \max\{|1 - \alpha\mu|, |1 - \alpha L|\}, \tag{4.16}$$

as in (3.18) given for quadratic functions. Using $\rho = \rho_{GD}(\alpha)$ in (4.14) leads to the following upper bound for GD.[6]

COROLLARY 4.4. *Let $f \in S_{\mu,L}(\mathbb{R}^d)$. Consider the GD iterations given by (4.12) with constant stepsize $\alpha \in (0, 2/L)$. Then, for all $k \geq 0$,*

$$\mathbb{E}[f(x_k) - f^*] \leq \rho_{GD}(\alpha)^{2k}\ \psi_0 + \left(1 - \rho_{GD}(\alpha)^{2k}\right)\ \sigma^2 R_{GD}(\alpha),$$

where $\psi_0 = \frac{L}{2}\|x_0 - x^*\|^2$ and $\rho_{GD}(\alpha)$ is given in (4.16). As a consequence,

$$\mathcal{J} \leq R_{GD}(\alpha), \quad \text{where} \quad R_{GD}(\alpha) \triangleq \frac{L\alpha^2 d}{2(1 - \rho_{GD}^2(\alpha))}. \tag{4.17}$$

Proof. Using the fact that $f(x_k) - f(x^*) \leq \frac{L}{2}\|x_k - x^*\|^2$ for $k \geq 0$ together with Proposition 4.3 yields the desired result. ∎

Note that by substituting ρ_{GD} in (4.17), we obtain $R_{GD} = \mathcal{O}(\alpha d)$. This bound is tight, as Proposition 3.2 implies that for quadratic functions $\mathcal{J} = \Theta(\alpha d)$.

4.4. Accelerated gradient method for strongly convex functions. We next consider the AG algorithm with gradient noise given by

$$x_{k+1} = y_k - \alpha(\nabla f(y_k) + w_k), \tag{4.18a}$$

$$y_k = (1+\beta)x_k - \beta x_{k-1}. \tag{4.18b}$$

As before, these iterations admit the dynamical system representation in (4.1) with A, B, C as in (2.6). We use the following result, which extends Lemma 3 in [28] to the case with noisy gradient.

[6] Trivially, $\mathcal{J}' \leq \frac{\alpha^2 d}{1-\rho_{GD}(\alpha)^2}$. More details are provided in Appendix E as supplementary material.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 05/03/20 to 18.9.61.111. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

LEMMA 4.5. *Let $f \in S_{\mu,L}(\mathbb{R}^d)$, and consider the dynamical system representation of AG with $\xi_k = [x_k^\top, x_{k-1}^\top]^\top$. Then, for any $\rho \in (0,1)$,*

$$\begin{bmatrix}\xi_k - \xi^* \\ \nabla f(y_k)\end{bmatrix}^\top \left(\rho^2 X_1 + (1-\rho^2)X_2\right) \begin{bmatrix}\xi_k - \xi^* \\ \nabla f(y_k)\end{bmatrix} \le \rho^2(f(x_k) - f^*) - (f(x_{k+1}) - f^*) + \frac{L\alpha^2}{2}\|w_k\|^2 - \alpha(1-L\alpha)\nabla f(y_k)^\top w_k$$

holds for all $k \ge 0$, where $X_1 = \tilde{X}_1 \otimes I_d$ and $X_2 = \tilde{X}_2 \otimes I_d$ with

$$\tilde{X}_1 = \frac{1}{2}\begin{bmatrix}\beta^2\mu & -\beta^2\mu & -\beta \\ -\beta^2\mu & \beta^2\mu & \beta \\ -\beta & \beta & \alpha(2-L\alpha)\end{bmatrix},\ \tilde{X}_2 = \frac{1}{2}\begin{bmatrix}(1+\beta)^2\mu & -\beta(1+\beta)\mu & -(1+\beta) \\ -\beta(1+\beta)\mu & \beta^2\mu & \beta \\ -(1+\beta) & \beta & \alpha(2-L\alpha)\end{bmatrix}.$$

Proof. Setting $x = x_k$ and $y = y_k$ in the second inequality in (1.1) leads to

$$f(x_k) - f(y_k) \ge \nabla f(y_k)^\top (x_k - y_k) + \frac{\mu}{2}\|x_k - y_k\|^2. \tag{4.19}$$

Similarly, setting $x = x_{k+1} = y_k - \alpha\nabla f(y_k) - \alpha w_k$ and $y = y_k$ in (1.1) yields

$$\begin{aligned} f(y_k) - f(x_{k+1}) &\ge \nabla f(y_k)^\top(\alpha\nabla f(y_k) + \alpha w_k) - \frac{L\alpha^2}{2}\|\nabla f(y_k) + w_k\|^2 \\ &= \frac{\alpha}{2}(2-L\alpha)\|\nabla f(y_k)\|^2 - \frac{L\alpha^2}{2}\|w_k\|^2 + \alpha(1-L\alpha)\nabla f(y_k)^\top w_k. \end{aligned} \tag{4.20}$$

Summing up (4.19) and (4.20) implies

$$\begin{aligned} f(x_k) - f(x_{k+1}) \ge \frac{1}{2}\begin{bmatrix}x_k - y_k \\ \nabla f(y_k)\end{bmatrix}^\top \begin{bmatrix}\mu I_d & I_d \\ I_d & \alpha(2-L\alpha)I_d\end{bmatrix}\begin{bmatrix}x_k - y_k \\ \nabla f(y_k)\end{bmatrix} \\ - \frac{L\alpha^2}{2}\|w_k\|^2 + \alpha(1-L\alpha)\nabla f(y_k)^\top w_k. \end{aligned} \tag{4.21}$$

Note that $x_k - y_k = x_k - ((1+\beta)x_k - \beta x_{k-1}) = \beta(x_{k-1} - x_k)$; hence, (4.21) implies

$$f(x_k) - f(x_{k+1}) + \frac{L\alpha^2}{2}\|w_k\|^2 - \alpha(1-L\alpha)\nabla f(y_k)^\top w_k \ge \begin{bmatrix}x_{k-1} - x^* \\ x_k - x^* \\ \nabla f(y_k)\end{bmatrix}^\top X_1 \begin{bmatrix}x_{k-1} - x^* \\ x_k - x^* \\ \nabla f(y_k)\end{bmatrix}. \tag{4.22}$$

Next, in a similar way, setting $x = x^*$ and $y = y_k$ in (1.1), and summing the second inequality with (4.20) leads to

$$f(x^*) - f(x_{k+1}) + \frac{L\alpha^2}{2}\|w_k\|^2 - \alpha(1-L\alpha)\nabla f(y_k)^\top w_k \ge \begin{bmatrix}x_{k-1} - x^* \\ x_k - x^* \\ \nabla f(y_k)\end{bmatrix}^\top X_2 \begin{bmatrix}x_{k-1} - x^* \\ x_k - x^* \\ \nabla f(y_k)\end{bmatrix}. \tag{4.23}$$

Multiplying (4.22) by ρ^2 and (4.23) by $1-\rho^2$, and summing them will lead to the desired result. ∎

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 05/03/20 to 18.9.61.111. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

PROPOSITION 4.6. *Let $f \in S_{\mu,L}(\mathbb{R}^d)$, and consider the AG iterations given by (4.18). Assume there exist $\rho \in (0,1)$, $P \in \mathbb{S}^{2d}_+$, and $c_0, c \geq 0$ such that*

$$c_0 X_0 + cX(\rho) \succeq \Phi(A, B, P, \rho), \quad \text{where} \tag{4.24}$$

$$X_0 = \begin{bmatrix} 2\mu L\, C^\top C & -(\mu+L)C^\top \\ -(\mu+L)C & 2I_d \end{bmatrix}, \quad X(\rho) = \rho^2 X_1 + (1-\rho^2)X_2$$

for X_1 and X_2 defined in Lemma 4.5. Then the following bounds hold for all $k \geq 0$:

$$\mathbb{E}[V_{P,c}(\xi_k)] \leq \rho^{2k} V_{P,c}(\xi_0) + \frac{1-\rho^{2k}}{1-\rho^2}\sigma^2\alpha^2\left(\frac{c}{2}Ld + \operatorname{Tr}(P_{11})\right), \tag{4.25}$$

where $P_{11} \in \mathbb{S}^d_+$ is the submatrix of P formed by its first d rows and d columns.

Proof. Using (1.2) for $x = y_k$ and $y = x^*$ along with the fact $y_k = C\xi_k$ yields

$$\begin{bmatrix} \xi_k - \xi^* \\ \nabla f(y_k) \end{bmatrix}^\top X_0 \begin{bmatrix} \xi_k - \xi^* \\ \nabla f(y_k) \end{bmatrix} \leq 0. \tag{4.26}$$

This inequality along with Lemma 4.5 implies that (4.10) holds for $X = c_0 X_0 + cX(\rho)$ and $\Gamma = \frac{1}{2}L\alpha^2 d$. Moreover, (4.24) implies that (4.9) holds for this X. Therefore, (4.11) holds and $\operatorname{Tr}(B^\top P B) = \alpha^2 \operatorname{Tr}(P_{11})$ completes the proof. ∎

As stated in the previous proposition, the MI in (4.24) provides us with (P, ρ) pairs and through (4.25) we obtain an upper bound on $\sup_k \mathbb{E}[V_{P,c}(\xi_k)]$, which leads to a bound R on $\mathcal{J}$. However, solving this $2d \times 2d$ MI becomes intractable as d increases. To keep this MI invariant of the dimension, we restrict our attention to the case that P is in the form of $\tilde{P} \otimes I_d$, where $\tilde{P}$ is a 2×2 symmetric positive semidefinite matrix; hence, (4.24) becomes a 3×3 MI. The following corollary shows the robustness bound when $P = \tilde{P} \otimes I_d$ for some $\tilde{P} \in \mathbb{S}^2_+$.

COROLLARY 4.7. *Let $f \in S_{\mu,L}(\mathbb{R}^d)$, and consider the AG iterations given by (4.18) with parameters α and β. Assume there exist $\rho \in (0,1)$, $\tilde{P} \in \mathbb{S}^2_+$, $c_0 \geq 0$, and $c > 0$ such that $c_0 X_0 + cX(\rho) \succeq \Phi(A, B, P, \rho)$ with X_0 defined in Theorem 4.6, X_1, X_2 defined in Lemma 4.5, and $P = \tilde{P} \otimes I_d$. Then for all $k \geq 0$,*

$$\mathbb{E}[f(x_k) - f^*] \leq \rho^{2k}\ \psi_0 + \left(1-\rho^{2k}\right)\sigma^2 R_{AG}(\alpha, \beta), \tag{4.27}$$

$$R_{AG}(\alpha,\beta) \triangleq \begin{cases} \dfrac{L\alpha^2 d}{2(1-\rho^2)}\ \dfrac{cL + 2\tilde{P}_{11}}{cL + 2(\tilde{P}_{11} - \tilde{P}_{12}^2/\tilde{P}_{22})}, & \tilde{P}_{22} > 0, \\ \dfrac{L\alpha^2 d}{2(1-\rho^2)}, & \tilde{P}_{22} = 0, \end{cases} \tag{4.28}$$

where $\psi_0 = \frac{1}{c}V_{P,c}(\xi_0)$. As a consequence, $\mathcal{J} \leq R_{AG}(\alpha, \beta)$.

Proof. Note that since $\tilde{P} \in \mathbb{S}^2_+$, we have $\tilde{P}_{22} \geq 0$. If $\tilde{P}_{22} > 0$, then using Schur complements, $\tilde{P}$ can be written as sum of two positive semidefinite matrices:

$$\tilde{P} = \begin{bmatrix} \tilde{P}_{11} - \tilde{P}_{12}^2/\tilde{P}_{22} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} \tilde{P}_{12}^2/\tilde{P}_{22} & \tilde{P}_{12} \\ \tilde{P}_{12} & \tilde{P}_{22} \end{bmatrix}.$$

Hence, $V_P(\xi_k) \geq (\tilde{P}_{11} - \tilde{P}_{12}^2/\tilde{P}_{22})\|x_k - x^*\|^2$. On the other hand, if $\tilde{P}_{22} = 0$, then $\tilde{P} \in \mathbb{S}^2_+$ implies that $\tilde{P}_{12} = 0$ as well and we get $V_P(\xi_k) \geq \tilde{P}_{11}\|x_k - x^*\|^2$. In either case, substituting the derived lower bounds on $V_P(\xi_k)$ in (4.25) and using the facts that $\frac{2}{L}(f(x_k) - f^*) \leq \|x_k - x^*\|^2$ and $\operatorname{Tr}(B^\top P B) = \operatorname{Tr}(BB^\top P) = \alpha^2 \tilde{P}_{11} d$ completes the proof. ∎

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 05/03/20 to 18.9.61.111. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Remark 4.1. Note that for $\alpha \in (0, \frac{2}{L})$, setting $\beta = 0$ in AG yields GD algorithm with stepsize α. Selecting $c_0 = 1$, $c = 0$, and $\tilde{P} = [\begin{smallmatrix} \tilde{p} & 0 \\ 0 & 0 \end{smallmatrix}]$, we observe that $\tilde{p} = L^2$ satisfies (4.24); therefore, Corollary 4.7 implies that $\mathcal{J} \leq \frac{L\alpha^2 d}{2(1-\rho_{GD}(\alpha)^2)}$ for GD, and hence, the result in (4.17) can be derived as a special case of Corollary 4.7.

Using this result, the next corollary characterizes the rate and robustness of the AG method with a particular parameterization.

COROLLARY 4.8. *Letting $f \in S_{\mu,L}(\mathbb{R}^d)$, consider the AG iterations given by* (4.18) *with constant stepsize $\alpha \in (0, 1/L]$ and $\beta(\alpha) = \frac{1-\sqrt{\alpha\mu}}{1+\sqrt{\alpha\mu}}$. Then, for all $k \geq 0$,*

$$\mathbb{E}[f(x_k) - f^*] \leq \rho_{AG}(\alpha)^{2k}\ \psi_0 + \left(1 - \rho_{AG}(\alpha)^{2k}\right)\ \sigma^2 R_{AG}(\alpha),$$

where $\psi_0 = V_{P,1}(\xi_0)$, $\rho_{AG}(\alpha) \triangleq \sqrt{1-\sqrt{\alpha\mu}}$, and $R_{AG}(\alpha) \triangleq \frac{\alpha d}{2(1-\rho_{AG}(\alpha)^2)}(1+\alpha L)$; hence, $\mathcal{J} \leq R_{AG}(\alpha) = \frac{\sqrt{\alpha} d}{2\sqrt{\mu}}(1+\alpha L)$.

Proof. [1, Theorem 2.3] guarantees that for any $\alpha \in (0, 1/L]$, the MI $X(\rho_{AG}(\alpha)) \succeq \Phi(A, B, P(\alpha), \rho_{AG}(\alpha))$ holds for A and B as in (2.6) corresponding to given α and $\beta(\alpha) = \frac{1-\sqrt{\alpha\mu}}{1+\sqrt{\alpha\mu}}$, where $P(\alpha) = \tilde{P}(\alpha) \otimes I_d$ for

$$\tilde{P}(\alpha) \triangleq \begin{bmatrix} \sqrt{\frac{1}{2\alpha}} \\ \sqrt{\frac{\mu}{2}} - \sqrt{\frac{1}{2\alpha}} \end{bmatrix} \begin{bmatrix} \sqrt{\frac{1}{2\alpha}} & \sqrt{\frac{\mu}{2}} - \sqrt{\frac{1}{2\alpha}} \end{bmatrix}. \tag{4.29}$$

Therefore, the desired result follows from Corollary 4.7. ∎

It is worth noting that although solving $c_0 X_0 + c X(\rho) \succeq \Phi(A, B, P, \rho)$ considering only lower-dimensional $P = \tilde{P} \otimes I_d$ is more restrictive, this small-dimensional MI can still recover the well-known rate $\bar{\rho}_{AG} = \sqrt{1 - \sqrt{\frac{1}{\kappa}}}$ for the deterministic case, i.e., $\sigma = 0$, in the literature [37] by setting the stepsize $\alpha = \frac{1}{L}$ and momentum parameter $\beta(\alpha) = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$. As shown in [28], this claim can be verified by setting $P = \tilde{P}(\alpha) \otimes I_d$ with $\alpha = 1/L$. In addition, for the case $L = \mu$, substituting $\beta(\alpha)$ in the explicit expression of $\mathcal{J}$ in (3.34) for quadratic functions, we obtain $\mathcal{J}(\alpha, \beta(\alpha)) = \Theta(\frac{\sqrt{\alpha} d}{\sqrt{\mu}})$, which implies that R_{AG} is a tight bound for $\mathcal{J}$ in terms of α dependency.

It is worth noting that the best rate known in the literature for general $f \in S_{\mu,L}(\mathbb{R}^d)$ is $\rho_* = \sqrt{1 - \sqrt{2\kappa - 1}/\kappa}$ provided in [42, Theorem 7]. However, this rate differs from $\bar{\rho}_{AG}$ just by a constant factor, i.e., $\frac{1-\rho_*^2}{1-\bar{\rho}_{AG}^2} \leq \sqrt{2}$. Moreover, for the special case of $f \in S_{\mu,L}(\mathbb{R}^d)$ being a quadratic function, the best linear rate for AG is $1 - 2/\sqrt{3\kappa + 1}$ (see [33, Proposition 1] for an asymptotic analysis and [6] for a non-asymptotic analysis). Therefore, we can conclude that $\rho = \mathcal{O}(\sqrt{1 - 1/\sqrt{\kappa}})$ and $\rho = \mathcal{O}(1 - 1/\sqrt{\kappa})$ denote the best known κ dependency of the rate coefficient for general and quadratic $f \in S_{\mu,L}(\mathbb{R}^d)$, respectively. That said, since the focus of section 4 is on general strongly convex functions $f \in S_{\mu,L}(\mathbb{R}^d)$, in the following subsection in order to approximate the rate-robustness trade-off curves for AG, we consider $\bar{\rho}_{AG} = \sqrt{1 - 1/\sqrt{\kappa}}$ as the reference rate as it exhibits the optimal κ dependency.

4.5. Approximating the rate and robustness trade-off curve. Similar to (3.22), (3.23), and (3.24) in section 3, there are several ways of forming an optimization problem to trade off the rate and robustness. In this section we focus on strongly

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 05/03/20 to 18.9.61.111. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

convex objectives $f \in S_{\mu,L}(\mathbb{R}^d)$, where unlike the quadratic objectives, we have access to upper bounds for the robustness measure rather than exact expressions. Therefore, we adopt a formulation similar to (3.22) and vary ϵ to characterize the rate-robustness trade-off. In fact, the parameter ϵ shows how much we desire to lose in terms of convergence rate to gain robustness.

In the rest, let $\bar{\rho}_{GD} \triangleq 1 - \frac{2}{\kappa+1}$ and $\bar{\rho}_{AG} \triangleq \sqrt{1-\sqrt{1/\kappa}}$ denote the linear convergence rates for GD and AG from the literature for $f \in S_{\mu,L}(\mathbb{R}^d)$ [37]. In this section, we assume $\kappa \neq 1$, as the $\kappa = 1$ case is trivial. We also let $\mathcal{J}_{\mathrm{GD},\epsilon}$ and $\mathcal{J}_{\mathrm{AG},\epsilon}$ be the best robustness value GD and AG can achieve corresponding to rate $\rho_{\mathrm{GD},\epsilon} \triangleq (1+\epsilon)\bar{\rho}_{GD}$ and $\rho_{\mathrm{AG},\epsilon} \triangleq (1+\epsilon)\bar{\rho}_{AG}$, respectively. In the rest of this section, we discuss methodologies to derive tractable upper bounds on $\mathcal{J}_{\mathrm{GD},\epsilon}$ and $\mathcal{J}_{\mathrm{AG},\epsilon}$.

In particular, for GD, the best robustness level while asking for linear convergence with rate $\rho_{\mathrm{GD},\epsilon}$ or faster is obtained by solving

$$\underset{\alpha\in(0,2/L)}{\arg\min}\ R_{GD}(\alpha) \quad \text{subject to} \quad \rho_{GD}(\alpha) \le \rho_{\mathrm{GD},\epsilon}, \tag{4.30}$$

where $\rho_{GD}(\alpha)$ is given in (4.16). The function $\rho_{GD}(\alpha)$ is convex and piecewise linear in α over the interval $[0, 2/L]$ with a unique minimum at $\bar{\rho}_{GD}$ and it satisfies $\rho_{GD}(0) = \rho_{GD}(2/L) = 1$ on the boundary points. Therefore, it follows from this property that, given $\epsilon \in (0, \frac{2}{\kappa-1})$, there are exactly two $\alpha_\epsilon > 0$ values such that $\rho_{GD}(\alpha_\epsilon) = \rho_{\mathrm{GD},\epsilon}$ which we can explicitly compute as $\alpha_\epsilon = \frac{2-\epsilon(\kappa-1)}{L+\mu}$ or $\alpha_\epsilon = \frac{2+\epsilon(\kappa-1)/\kappa}{L+\mu}$. The former value is strictly smaller as $\varepsilon > 0$ and $\kappa > 1$ here. From the formula (4.17), we have $R_{GD}(\alpha_\varepsilon) = \frac{L\alpha_\varepsilon^2 d}{2(1-\rho^2_{\mathrm{GD},\epsilon})}$. Clearly one should select the smaller α_ϵ value to minimize the robustness bound, i.e., a choice of $\alpha_\epsilon = \frac{2-\epsilon(\kappa-1)}{L+\mu}$ leads to $\rho_{\mathrm{GD},\epsilon}$ rate with a robustness bound $R_{GD}(\alpha_\epsilon)$, i.e., $\mathcal{J}_{\mathrm{GD},\epsilon} \le R_{GD}(\alpha_\epsilon)$.

For AG, we can also write an analogous optimization problem in order to trade rate with robustness:

$$\underset{\substack{\alpha,\beta\ge 0,\tilde{P}\in\mathbb{S}^2_+\\ \rho,c_0,c\ge 0}}{\arg\min}\ \{R_{AG}(\alpha,\beta):\ \rho \le \rho_{\mathrm{AG},\epsilon},\ c_0 X_0 + cX(\rho) \succeq \Phi(A,B,\tilde{P}\otimes I_d,\rho)\} \tag{4.31}$$

with $X_0, X(\rho), R_{AG}$ defined in Corollary 4.7—since we can scale $\tilde{P}$ with $c > 0$, without loss of generality, we restrict our attention to the case $c = 1$ for treating the $c > 0$ case. This problem is in general nonconvex and not easy to solve. Here, we consider two different ways to generate rate-robustness trade-off curves.

The first approach is similar to the one we used for GD. In particular, consider Corollary 4.8, for $\alpha \in (0, 1/L]$, choosing $\beta = \frac{1-\sqrt{\alpha\mu}}{1+\sqrt{\alpha\mu}}$ implies that $\rho_{AG}(\alpha) = \sqrt{1-\sqrt{\alpha\mu}}$. We get $\rho_{AG}(\alpha_\epsilon) = \rho_{\mathrm{AG},\epsilon}$ for

$$\alpha_\epsilon = [1-\rho^2_{\mathrm{AG},\epsilon}]^2/\mu = \left[1-(1+\epsilon)^2\left(1-\frac{1}{\sqrt{\kappa}}\right)\right]^2 \Big/ \mu, \tag{4.32}$$

with $\epsilon \in \left[0, \sqrt{\frac{\sqrt{\kappa}}{\sqrt{\kappa}-1}} - 1\right)$ to make sure the rate is smaller than 1. Thus, choosing $(\alpha,\beta) = (\alpha_\epsilon, \beta_\epsilon)$ with $\beta_\epsilon \triangleq \frac{1-\sqrt{\alpha_\epsilon\mu}}{1+\sqrt{\alpha_\epsilon\mu}}$ guarantees the rate $\rho_{\mathrm{AG},\epsilon}$. In addition, Corollary 4.8 implies the robustness bound $R_{AG}(\alpha_\epsilon) = \mathcal{O}(\frac{\sqrt{\alpha_\epsilon}d}{\sqrt{\mu}})$ for this case.

For the second approach, we first grid the $(\alpha,\beta) \in (0, \frac{2}{L}] \times [0,1]$ parameter space and use a numerical approach to find the best parameters for each ϵ, i.e., we solve

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 05/03/20 to 18.9.61.111. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

a low-dimensional (in $\mathbb{R}^4$) convex semidefinite programming (SDP) problem for each possible (α, β) value from the grid, and we will pick the best one to determine the robustness bound. The SDPs arise from a convex approximation to the problem (4.31) as we now elaborate further. First, we note that $R_{AG}(\alpha, \beta)$ defined in (4.28) is not convex in $\tilde{P}$; therefore, to obtain a tractable problem, we replace $R_{AG}(\alpha, \beta)$ with a convex upper bound. In particular, using Proposition 4.6, it is straightforward to see

$$R_{AG}(\alpha, \beta) \leq \bar{R}_{AG}(\alpha, \beta) \triangleq \frac{\alpha^2 d}{2(1-\rho^2)}(L + 2\tilde{P}_{11})$$

whenever there exists $\rho \in (0,1)$, $\tilde{P} \in \mathbb{S}^2_+$ and $\bar{c} \geq 0$ such that[7]

$$\bar{c}X_0 + X(\rho) \succeq \Phi(A, B, \tilde{P} \otimes I_d, \rho). \tag{4.33}$$

Note given $\epsilon \in [0, \sqrt{\frac{\sqrt{\kappa}}{\sqrt{\kappa}-1}} - 1)$, setting $\rho = \rho_{\mathrm{AG},\epsilon}$ within the MI in (4.33), we get a linear MI in $\bar{c} \geq 0$ and $\tilde{P} \in \mathbb{S}^2_+$ for fixed (α, β). Moreover, $\bar{R}_{AG}(\alpha, \beta)$ is linear in $\tilde{P}$. Hence, given the trade-off parameter $\epsilon > 0$, we will approximately solve

$$\bar{\mathcal{R}}(\epsilon) \triangleq \min_{\alpha,\beta,\bar{c}\geq 0, \tilde{P}\in\mathbb{S}^2_+} \{\bar{R}_{AG}(\alpha, \beta) : \ \bar{c}X_0 + X(\rho_{\mathrm{AG},\epsilon}) \succeq \Phi(A, B, \tilde{P} \otimes I_d, \rho_{\mathrm{AG},\epsilon})\} \tag{4.34}$$

with X_0 and $X(\rho)$ defined in Corollary 4.7, and $\bar{R}_{AG}$ as given above. In fact, for a fixed (α, β), this is a small dimensional convex SDP problem and can be solved easily.

Thus, we first grid the AG parameter space, i.e., $\{(\alpha_{i_1}, \beta_{i_2})\}_{i_1\in\mathcal{I}_1, i_2\in\mathcal{I}_2}$, and for given trade-off parameter ϵ, we solve $|\mathcal{I}_1||\mathcal{I}_2|$ many four-dimensional SDPs, i.e., for each $(i_1, i_2) \in \mathcal{I}_1 \times \mathcal{I}_2$,

$$\begin{aligned} \bar{\mathcal{R}}_{i_1,i_2}(\epsilon) \triangleq \min_{\bar{c}\geq 0, \tilde{P}\in\mathbb{S}^2_+} \bar{R}_{AG}(\alpha_{i_1}, \beta_{i_2}) &= \frac{\alpha_{i_1}^2 d}{2(1-\rho^2_{\mathrm{AG},\epsilon})}(L + 2\tilde{P}_{11}) \\ \text{s.t.} \quad \bar{c}X_0 + X(\rho_{\mathrm{AG},\epsilon}) &\succeq \Phi(A, B, \tilde{P} \otimes I_d, \rho_{\mathrm{AG},\epsilon}). \end{aligned} \tag{4.35}$$

Clearly, $\mathcal{J}_{\mathrm{AG},\epsilon} \leq \bar{\mathcal{R}}(\epsilon) \leq \min_{i_1\in\mathcal{I}_1, i_2\in\mathcal{I}_2} \bar{\mathcal{R}}_{i_1,i_2}(\epsilon)$, where $\bar{\mathcal{R}}(\epsilon)$ is as in (4.34). Note $(\alpha, \beta) = (\alpha_\epsilon, \beta_\epsilon)$, $\bar{c} = 0$, and $\tilde{P} = \tilde{P}(\alpha_\epsilon)$ satisfies (4.33) where α_ϵ is given in (4.32), $\beta_\epsilon = \frac{1-\sqrt{\alpha_\epsilon \mu}}{1+\sqrt{\alpha_\epsilon \mu}}$ and $\tilde{P}(\alpha)$ is defined in (4.29). Therefore, for any grid that contains $(\alpha_\epsilon, \beta_\epsilon)$ as one of the grid points, we have

$$\bar{\mathcal{R}}(\epsilon) \leq \frac{\alpha_\epsilon^2 d}{2\left(1-\rho^2_{\mathrm{AG},\epsilon}\right)}\left(L + \frac{1}{\alpha_\epsilon}\right) \leq \frac{\alpha_\epsilon d}{\left(1-\rho^2_{\mathrm{AG},\epsilon}\right)} = \frac{\sqrt{\alpha_\epsilon} d}{\sqrt{\mu}}, \tag{4.36}$$

where the first inequality follows from $\tilde{P}_{11} = \frac{1}{2\alpha_\epsilon}$ for $\tilde{P} = \tilde{P}(\alpha_\epsilon)$ and the second inequality follows from $\alpha_\epsilon \in (0, 1/L]$. The equality above is a direct consequence of the identity (4.32). Finally, according to the discussion at the end of the previous section, for $\epsilon > 0$ sufficiently close to $\sqrt{\frac{\sqrt{\kappa}}{\sqrt{\kappa}-1}} - 1$, there exists a quadratic $f \in \mathbb{S}_{\mu,L}$ such that $\mathcal{J}(\alpha_\epsilon, \beta_\epsilon) = \Theta(\frac{\sqrt{\alpha_\epsilon} d}{\sqrt{\mu}})$, while (4.36) implies that $\mathcal{J}_{\mathrm{AG},\epsilon} \leq \bar{\mathcal{R}}(\epsilon) \leq \frac{\sqrt{\alpha_\epsilon} d}{\sqrt{\mu}}$.

In Figure 3, we consider the rate-robustness trade-off for a strongly convex function with $\mu = 1$ and $L = 20$ using the upper bounds we derived in this section for both GD and AG (which are applicable to any dimension d), where we report the normalized robustness level $\mathcal{J}/d$ in the y-axis versus the convergence rate in the x-axis—for

[7] Recall that we put $c = 1$ in this section.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

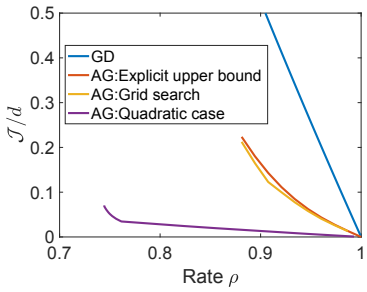

FIG. 3. *Rate-robustness trade-off for GD and AG algorithms using derived upper bounds and comparing it with the quadratic result.*

AG we plotted two curves associated with the two approaches detailed above: the first one (marked in red) uses the explicit bound and the second one (marked in yellow) is based on a grid search. We use the solver CVX [25] to solve the four-dimensional SDPs given in (4.35). We select a grid of size $|I_1| = |I_2| = 30$ on the parameter space for (α, β) determined by Proposition 3.4. We observe that the grid search and the explicit upper bound approach give similar results on this numerical example for AG, especially when the rate is close to 1. We also see that the robustness upper bound for GD obtained from our approach (marked in blue) is worse than the upper bound we developed for AG. To see how the AG bounds are comparable with the exact expressions we developed for quadratics, first we note that by Lemma D.1, any quadratic function $f \in S_{\mu,L}(\mathbb{R}^d)$ admits the robustness upper bound[8]

$$\mathcal{J}_{max}(\alpha,\beta) := (d-1)\max\left[u_{\alpha,\beta}(\mu), u_{\alpha,\beta}(L)\right] + \min\left[u_{\alpha,\beta}(\mu), u_{\alpha,\beta}(L)\right]$$

and this bound can be achieved for some choices of f. For large d, we have clearly $\mathcal{J}_{max}(\alpha,\beta)/d \approx \max\left[u_{\alpha,\beta}(\mu), u_{\alpha,\beta}(L)\right]$. In Figure 3, we plot the latter quantity versus the convergence rate (marked in purple) to demonstrate the rate-robustness curve for AG in the case of quadratic objective functions. We observe from Figure 3 that our bounds for the quadratic case are tighter than those for general strongly convex functions as expected.

5. Asymptotic stability of the optimum with respect to perturbations. Our discussion so far has focused on the robustness of first-order methods with respect to random noise in the gradients, which we quantify by $\mathcal{J}$ defined in (2.8). Our robustness measure $\mathcal{J}$ is based on the H_2 norm of an associated linear dynamical system. It is well-known that the H_2 norm of a dynamical system is closely related to the *asymptotic stability* of the equilibrium (which is characterized by the optimal solution x^* to (2.1) in our setup) in the sense that it quantifies how quickly the system can converge back to the equilibrium if it is unsettled from its equilibrium in the direction of a coordinate [46]. More specifically, for each $i \in \{1, \dots, d\}$, let $\{x_k^i\}_{k\geq 0}$ be the iterate sequence corresponding to (4.1) whenever $\{w_k\}_k = \delta[k]e_i$ for $k \geq 0$, where e_i is the ith basis vector, i.e., we perturb the system from its equilibrium with an impulse input in the direction of e_i. Let

[8] This is a worst-case bound for a quadratic $f \in S_{\mu,L}(\mathbb{R}^d)$; tighter bounds are available if all the eigenvalues of its Hessian matrix are known or estimated beyond the eigenvalues μ and L (see Proposition 3.5).

Downloaded 05/03/20 to 18.9.61.111. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

$$\mathcal{J}_* := \sum_{i=1}^{d} \|x^i - x^*\|_2^2, \tag{5.1}$$

where $\|x^i - x^*\|_2$ is the l_2 norm of the sequence $\{x_k^i - x^*\}_k$. It is worth noting that $\{x_k^i\}_k$ is the same as the iterate sequence of the noiseless system (2.2) with initial state $\xi^* + Be_i$, $D = 0_d$ and $\phi(\cdot) = \nabla f(\cdot)$.

We next motivate this definition from another perspective. Recall the alternative robustness measure $\mathcal{J}' = \limsup_k \mathbb{E}[\|x_k - x^*\|^2]$ we briefly discussed in section 2; for a linear system it is known that $\mathcal{J}' = \mathcal{J}_*$ [46]. In other words, our explicit formulas and bounds for $\mathcal{J}'$, given in Appendix E, translate immediately to $\mathcal{J}_*$ for optimizing *quadratic* functions. The definition in (5.1) also extends to the case when f is not necessarily quadratic or equivalently when the system (4.1) is nonlinear; however, for nonlinear systems there is no known explicit formula that relates $\mathcal{J}'$ to $\mathcal{J}_*$ [44]. We refer to the quantity $\mathcal{J}_*$ as *perturbation stability* of the first-order algorithm in consideration; indeed, it measures how sensitive the underlying optimization algorithm is to the initialization around the optimal solution—it also quantifies how strongly the iterate sequence is attracted to the optimal solution once they are close.

In particular, given A, B, and C defining the first-order optimization algorithm, we assume there exists $X \in \mathbb{S}^{m+d}$ such that the MI in (4.9) holds for some $P \in \mathbb{S}_+^m$ and $\rho \in (0,1)$. Moreover, we assume that when $\Gamma = 0$ and $\sigma = 0$, there exists a constant $c \geq 0$ independent of ξ_0 such that the same ρ and X satisfy the dissipation inequality in (4.10) for every $k \geq 0$; hence, it follows from (4.9) and (4.10) along with (4.7) that

$$V_{P,c}(\xi_k) \leq \rho^{2k} V_{P,c}(\xi_0) \quad \forall\, k \geq 0. \tag{5.2}$$

Since $f \in S_{\mu,L}(\mathbb{R}^d)$, whenever $P \in \mathbb{S}_+^m$ and/or $c > 0$, (5.2) implies that the error signal $\{\|x_k^i - x_*\|^2\}_k$ decays geometrically and is therefore summable, and as a consequence, $\mathcal{J}_*$ in (5.1) well-defined.

LEMMA 5.1. *For GD, the following bound holds for all $\alpha \in (0, 2/L)$:*

$$\mathcal{J}_*(\alpha) \leq \frac{\alpha^2 d}{1 - \rho_{GD}^2(\alpha)}, \tag{5.3}$$

where $\rho_{GD}(\cdot)$ is defined in (4.16). *Moreover, for AG, given $\alpha \in (0, 1/L]$, setting $\beta(\alpha) = \frac{1-\sqrt{\alpha\mu}}{1+\sqrt{\alpha\mu}}$, the perturbation stability can be bounded as $\mathcal{J}_*(\alpha) \leq \frac{\alpha^2 d}{\sqrt{\alpha\mu}}(1+\kappa)$.*

Proof. Recall that $\{x_k^i\}_k$ is the same as the iterate sequence of the noiseless system (2.2) with initial state $\xi^* + Be_i$. Hence, Proposition 4.3 with $\sigma = 0$ implies that

$$\left\|x_k^i - x^*\right\|^2 \leq \rho^{2k} \left\|x_0^i - x^*\right\|^2 \quad \forall\, k \geq 0 \tag{5.4}$$

for some $\rho \in (0,1)$ and for any $1 \leq i \leq d$ and stepsize $\alpha \in (0, 2/L)$. Thus, $\sum_{k=0}^{\infty} \left\|x_k^i - x^*\right\|^2 \leq \frac{1}{1-\rho^2} \left\|x_0^i - x^*\right\|^2$, which implies $\left\|x^i - x^*\right\|^2 \leq \frac{\alpha^2}{1-\rho^2}$ for all $i = 1, \ldots, d$ since $B = -\alpha I_d$ for GD. Therefore, we have $\mathcal{J}_*(\alpha) \leq \frac{\alpha^2 d}{1-\rho^2}$. Moreover, given any stepsize $\alpha \in (0, 2/L)$ for GD, using (4.16), which is the smallest ρ value for which (5.4) holds, we obtain (5.3). On the other hand, for AG, using Corollary 4.8 with $\sigma = 0$ and the fact that $\frac{\mu}{2} \|x_k - x^*\|^2 \leq f(x^k) - f^*$, we get $\left\|x_k^i - x^*\right\|^2 \leq \rho_{AG}^{2k}(\left\|x_0^i - x^*\right\|^2 + \frac{2}{\mu}(f(x_0^i) - f^*))$ for $k \geq 0$ and $i = 1, \ldots, d$, where we used $x_0^i = x_{-1}^i = x^* + Be_i$ for $i = 1, \ldots, d$. Thus,

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 05/03/20 to 18.9.61.111. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

$$\mathcal{J}_*(\alpha) \le \frac{1}{1-\rho_{AG}^2(\alpha)}\left(\mathrm{Tr}(B^\top B) + \sum_{i=1}^{d} \tfrac{2}{\mu}(f(x_0^i) - f^*)\right)$$
$$\le \frac{\alpha^2 d}{\sqrt{\alpha\mu}}(1+\kappa) \quad \forall\ \alpha \in (0, 1/L]. \qquad \square$$

For $\alpha \in (0, 1/L]$ since $\rho_{GD}(\alpha) = 1 - \alpha\mu$, (5.3) implies that $\mathcal{J}_*(\alpha) \le \frac{\alpha^2 d}{\alpha\mu(2-\alpha\mu)} = \mathcal{O}(\alpha)$ for GD. For quadratic $f \in S_{\mu,L}(\mathbb{R}^d)$ such that $\mu = L$, as we discussed in footnote 6, $J'(\alpha) = \frac{\alpha d}{\mu(2-\alpha\mu)}$. Since $J_*(\alpha) = J'(\alpha)$ for quadratics, $\mathcal{O}(\alpha)$ dependence of (5.3) for $\alpha \in (0, 1/L]$ is tight. We also note that $\mathcal{O}(\alpha^{3/2})$ bound on $\mathcal{J}_*(\alpha)$ for AG implies that for sufficiently small stepsize α, AG possesses better perturbation stability than GD.

6. Numerical experiments. Our first set of experiments concerns a further study of Example 3.2 for comparing AG and GD in terms of performance. In the leftmost plot of Figure 4, we vary the trade-off parameter from $\tau = 0$ to $\tau = \infty$ for AG and plot the robustness level $\mathcal{J}(\alpha_*(\tau), \beta_*(\tau))$ versus the rate $\rho(\alpha_*(\tau), \beta_*(\tau))$ corresponding to the optimal parameters $(\alpha_*(\tau), \beta_*(\tau))$, we also plot the analogous curve for GD (the same curve from Figure 1) to compare. We observe that for the same achievable convergence rate, the optimized AG parameters lead to more robust algorithms compared to the optimized GD algorithms as AG has an additional parameter β to optimize robustness over. This shows that AG can improve GD in terms of both convergence rate and robustness at the same time when gradients are subject to white noise. This result is in contrast with the deterministic gradient error setting in [13], which shows that GD performance degrades gracefully while AG may accumulate error. Therefore, our results suggest that AG algorithms can tolerate random noise better than deterministic noise to preserve their accelerated rates, which is also in line with the theoretical findings of [7]. Also, it is interesting to note that the popular choice of parameters (blue and red dots), as well as the parameters that lead to the optimal (fastest) rate (green and purple dots), lie on curves that trade robustness with rate in an optimal fashion.

Next, we illustrate the tightness of our upper bound $\bar{\mathcal{J}}(\alpha, \beta)$ provided in (3.36) to the (true) robustness level $\mathcal{J}(\alpha, \beta)$. This upper bound results in a small-scale optimization problem (3.37) that allows trading off robustness and the convergence rate in a way that is computationally tractable, even in high dimensions. The middle plot of Figure 4 shows the convergence rate and robustness obtained by solving (3.29) versus solving (3.37). The objective is a random quadratic function in dimension $d = 100$ with parameters $\mu = 0.1, L = 1$. Our results show that for any trade-off parameter τ our upper bound is within a factor of 1.2 of true parameters, illustrating the accuracy

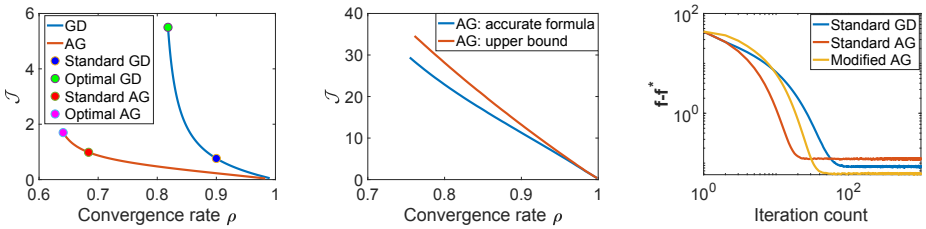

FIG. 4. *Left: Robustness $\mathcal{J}$ as a function of the convergence rate for GD and AG. Middle: Comparison of the convergence rate and robustness obtained by solving* (3.29) *versus its approximation* (3.37) *for* $d = 100$. *Right: Tuned AG can be both faster and more robust than GD.*

Downloaded 05/03/20 to 18.9.61.111. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

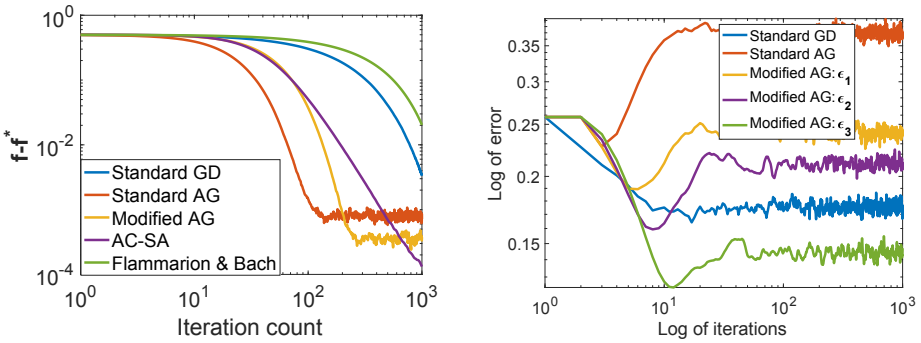

FIG. 5. *Left: Comparison of the other algorithms with modified AG. Right: Tuned AG for the regularized logistic regression problem.*

of this approximation to the optimal parameters for different levels of robustness, and especially, the approximation is more accurate when the trade-off parameter is larger (in which case the convergence rate is closer to 1). We obtain quantitatively similar results repeating this experiment with other randomly generated quadratic functions.

Next, we illustrate our framework to trade off robustness and convergence rate on a quadratic optimization problem, similar to the one considered in [27], where it is shown that AG algorithms with standard choice of parameters have difficulty handling random gradient noise. We consider the quadratic function $f(x) = \frac{1}{2}x^\top Qx + b^T x + \delta\|x\|^2$ in dimension $d = 100$, where Q is the Laplacian of a cyclic graph, $\delta = 0.1$ is a regularization parameter to make the problem strongly convex, and b is a random vector. As can be seen in the rightmost plot of Figure 4, we show that when properly modified, AG can be both faster and more robust in comparison with GD.

In the leftmost plot of Figure 5, we compare the tuned AG with other algorithms such as AC-SA [32] and the Flammarion–Bach algorithm [19]. For this purpose, we consider the same quadratic test problem from [19] in dimension $d = 20$, where the eigenvalues of its Hessian Q are set equal to $\lambda_i = i^2$ for $i = 1, 2, \ldots, 20$. Our results show that modified AG can trade robustness with the convergence rate successfully and can improve upon AC-SA and the Flammarion–Bach algorithm on this example.

Finally, we validate our results for strongly convex and smooth functions by choosing function f to be a regularized logistic loss. We synthesize a random matrix $M \in \mathbb{R}^{2000\times 100}$ and a random vector $w \in \mathbb{R}^{100}$ and compute $y = \text{sign}(Mw)$ as the output of the classifier. The goal is to recover w using regularized logistic regression when the gradient of the logistic loss is corrupted with additive Gaussian noise. The plot in the right panel of Figure 5 shows the behavior of tuned AG after solving the optimization problem (4.31) with three different ϵ values $0 < \epsilon_1 < \epsilon_2 < \epsilon_3$ in comparison with standard *AG* and *GD*. As predicted by our theory, AG performs better than GD and the asymptotic suboptimality decreases as ε gets larger.

7. Conclusion. We consider the gradient descent (GD) and accelerated gradient (AG) methods for optimizing strongly convex functions. We developed a computationally tractable framework to design their parameters in a way to trade between two conflicting performance measures: the convergence rate and the robustness to additive white noise in the gradient computations measured in terms of final asymptotic variance of the algorithm output. For strongly convex quadratics, we show that this robustness measure is equal to the H_2 norm of a dynamical system associated to

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 05/03/20 to 18.9.61.111. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Downloaded 05/03/20 to 18.9.61.111. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

the optimization algorithm and give an explicit characterization of this quantity. Our results show that for the same achievable rate, AG can always be tuned to be more robust. Similarly, for the same robustness level, we show that AG can be tuned to be always faster than GD. We also give fundamental lower bounds on the achievable robustness level for GD for a given achievable rate. We show how our analysis can be extended to smooth strongly convex functions and we derive upper bounds on the robustness measures for GD and AG.

Appendix A. First-order optimality conditions for the objective $F_\tau(\alpha)$.

PROPOSITION A.1. *There exists an optimizer $\alpha_*(\tau)$ to the minimization problem (3.24). Furthermore, any optimizer either is $\alpha_*(\tau) = 2/(\mu+L)$ or satisfies one of the following two conditions:*

$$\text{(A.1)} \qquad \begin{cases} \dfrac{\alpha^2}{2}\displaystyle\sum_{i=1}^{d} \dfrac{1}{(2-\alpha\lambda_i)^2} + \dfrac{\tau(\alpha\mu-1)}{\mu(2-\alpha\mu)^2} = 0 \quad \textit{and} \quad |1-\alpha\mu| > |1-\alpha L|, \end{cases}$$

$$\text{(A.2)} \qquad \begin{cases} \dfrac{\alpha^2}{2}\displaystyle\sum_{i=1}^{d} \dfrac{1}{(2-\alpha\lambda_i)^2} + \dfrac{\tau(\alpha L-1)}{L(2-\alpha L)^2} = 0 \quad \textit{and} \quad |1-\alpha\mu| < |1-\alpha L|. \end{cases}$$

Therefore, by examining the values of F at the points that satisfy these equality and inequality constraints, we can determine the optimal stepsize $\alpha^*(\tau)$.

Proof. To solve (3.24) explicitly, note that

$$\frac{1}{1-\rho^2} = \begin{cases} \frac{1}{1-(1-\alpha\mu)^2} & \text{if} \quad |1-\alpha\mu| > |1-\alpha L|, \\ \frac{1}{1-(1-\alpha L)^2} & \text{if} \quad |1-\alpha\mu| \le |1-\alpha L|. \end{cases}$$

The optimal α^* cannot be attained on the boundary points of the interval $[0, 2/L]$ as F is not finite at these points. Therefore, it suffices to solve the optimization problem over the open interval $(0, 2/L)$ where F is differentiable with respect to α except when $|1-\alpha\mu| = |1-\alpha L|$, i.e., when $\alpha = 2/(\mu+L)$. For $\alpha^* \neq 2/(\mu+L)$, we can write the first-order conditions of optimality $\frac{\partial F}{\partial \alpha} = 0$, which leads to (A.1) and (A.2). ∎

Appendix B. Proof of Proposition 3.4. In light of the formula (3.31) that characterizes $\rho(A_Q)$, the closure of the stability set $\mathcal{S}$ admits the representation $\mathcal{S} = \mathcal{S}_\mu \cap \mathcal{S}_L$ where for $\lambda \in \{\mu, L\}$ we define

$$\text{(B.1)} \qquad \mathcal{S}_\lambda = \{(\alpha,\beta) \ : \ \rho_\lambda(\alpha,\beta) \le 1, \alpha \ge 0, \beta \ge 0\} \subset \mathbb{R}^2.$$

We first write $\mathcal{S}_\lambda$ as a union of two disjoint sets depending on the signature of Δ_λ: $\mathcal{S}_\lambda = \mathcal{S}_{\lambda,1} \cup \mathcal{S}_{\lambda,2}$, where

$$\text{(B.2)} \qquad \mathcal{S}_{\lambda,1} = \mathcal{S}_\lambda \cap \{(\alpha,\beta) \ : \ \Delta_\lambda \le 0\}, \quad \mathcal{S}_{\lambda,2} = \mathcal{S}_\lambda \cap \{(\alpha,\beta) \ : \ \Delta_\lambda > 0\}.$$

It follows from the definition of Δ_λ in (3.32) that $\Delta_\lambda \le 0$ if and only if $0 \le 1-\alpha\lambda \le \frac{4\beta}{(1+\beta)^2}$, and when this condition holds, $\rho_\lambda = \sqrt{\beta(1-\alpha\lambda)} \le 1$ if and only if $0 \le 1-\alpha\lambda \le \frac{1}{\beta}$. Therefore,

$$\text{(B.3)} \qquad \mathcal{S}_{\lambda,1} = \left\{ (\alpha,\beta) : \ 0 \le 1-\alpha\lambda \le \min\left\{ \frac{1}{\beta}, \frac{4\beta}{(1+\beta)^2} \right\} \right\}.$$

We next focus on $\mathcal{S}_{\lambda,2}$. Note that $\Delta_\lambda \ge 0$ if and only if

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 05/03/20 to 18.9.61.111. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

$$1-\alpha\lambda \le 0 \quad \text{or} \quad 1-\alpha\lambda \ge \frac{4\beta}{(1+\beta)^2}. \tag{B.4}$$

If (B.4) is satisfied, then $\rho_\lambda \le 1$ if and only if $\frac{1}{2}(1+\beta)(1-\alpha\lambda)\text{sign}(1-\alpha\lambda)+\frac{1}{2}\sqrt{\Delta_\lambda} \le 1$. There are two cases:

(1) $\Delta_\lambda > 0$ and $1-\alpha\lambda < 0$: In this case, $\rho_\lambda \le 1$ if and only if $\sqrt{\Delta_\lambda} \le 2-(1+\beta)c_\lambda$, where $c_\lambda = -(1-\alpha\lambda) > 0$. By squaring both sides, this is if and only if $\Delta_\lambda \le (2-(1+\beta)c_\lambda)^2$ and $2-(1+\beta)c_\lambda \ge 0$. The first inequality holds if $c_\lambda = -(1-\alpha\lambda) \le \frac{1}{2\beta+1}$, whereas the second inequality holds if $c_\lambda = -(1-\alpha\lambda) \le \frac{2}{\beta+1}$. The first inequality is more binding; if it holds, the second inequality holds too. Therefore,

$$\left\{(\alpha,\beta):\ 0 \le -(1-\alpha\lambda) \le \frac{1}{2\beta+1}\right\} \subset \mathcal{S}_{\lambda,2}. \tag{B.5}$$

(2) $\Delta_\lambda > 0$ and $1-\alpha\lambda > 0$: In this case, $\rho_\lambda \le 1$ if and only if $\sqrt{\Delta_\lambda} \le 2-(1+\beta)d_\lambda$, where $d_\lambda := -c_\lambda = (1-\alpha\lambda) > 0$. After squaring both sides, this is if and only if

$$\Delta_\lambda \le (2-(1+\beta)d_\lambda)^2 \quad \text{and} \quad 2-(1+\beta)d_\lambda \ge 0, \tag{B.6}$$

where the first inequality simplifies to $1 \ge d_\lambda$. (B.6) along with (B.4) means $\frac{4\beta}{(1+\beta)^2} \le 1-\alpha\lambda \le \min\{1, \frac{2}{1+\beta}\}$, which implies $\beta \le 1$; therefore,

$$\left\{(\alpha,\beta):\ \frac{4\beta}{(1+\beta)^2} \le 1-\alpha\lambda \le \frac{2}{1+\beta}\right\} \subset \mathcal{S}_{\lambda,2}. \tag{B.7}$$

Merging (B.3), (B.5), and (B.7) yields

$$S_\lambda = \left\{(\alpha,\beta):\ 1-\alpha\lambda \in \left[-\frac{1}{1+2\beta}, \min\left\{\frac{1}{\beta}, \frac{2}{1+\beta}\right\}\right]\right\}. \tag{B.8}$$

To complete the proof, due to the representation (B.1), it suffices to compute the intersection $\mathcal{S}_\mu \cap \mathcal{S}_L$. There are several cases to consider depending on the value of α:

(1) First, consider $\alpha \in [0, \frac{1}{L}]$. In this case $1-\alpha\mu \ge 1-\alpha L \ge 0$, and hence (B.8) implies $1-\alpha\mu \le \frac{2}{1+\beta}$ if $\beta \le 1$, whereas $1-\alpha\mu \le \frac{1}{\beta}$ if $\beta \ge 1$. Nevertheless, if $\beta \le 1$, then $\frac{2}{1+\beta} \ge 1$, so the first case always holds; hence, $(1-\alpha\mu)\beta \le 1$.

(2) Now, assume $\alpha \in [\frac{1}{L}, \min\{\frac{2}{L}, \frac{1}{\mu}\}]$. Then $1-\alpha\mu \ge 0 \ge 1-\alpha L$, and thus (B.8) yields $1-\alpha L \ge -\frac{1}{1+2\beta}$, $1-\alpha\mu \le \min\{\frac{1}{\beta}, \frac{2}{1+\beta}\}$, where the second inequality again simplifies to $(1-\alpha\mu)\beta \le 1$.

(3) The last possible case happens when $\mu \ge \frac{L}{2}$, and so $\alpha \in [\frac{1}{\mu}, \frac{2}{L}]$ is possible. In this case $1-\alpha L \le 1-\alpha\mu \le 0$, and so using (B.8), we just need to check $1-\alpha L \ge -\frac{1}{1+2\beta}$. Considering all these cases along with the fact that (B.8) shows α cannot be greater than $\frac{2}{L}$ completes the proof.

Appendix C. Proof of Proposition 3.5. Similar to the analysis for GD, we can assume without loss of generality that Q is diagonal. The proof is also similar. Consider $U\Lambda U^\top$ be the eigenvalue decomposition of Q. Then A_Q in (3.28) can be written as

$$A_Q = \tilde{U} A_\Lambda \tilde{U}^\top, \quad \text{where} \tag{C.1}$$

$$\tilde{U} = \begin{bmatrix} U & 0_d \\ 0_d & U \end{bmatrix}, A_\Lambda = \begin{bmatrix} (1+\beta)(I_d - \alpha\Lambda) & -\beta(I_d-\alpha\Lambda) \\ I_d & 0_d \end{bmatrix}. \tag{C.2}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 05/03/20 to 18.9.61.111. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Replacing A_Q from (C.1) in Lyapunov equation (3.9) implies

$$\tilde{U}A_\Lambda\tilde{U}^\top X\tilde{U}A_\Lambda^\top\tilde{U}^\top - X + BB^\top = 0. \tag{C.3}$$

Multiplying by $\tilde{U}$ and $\tilde{U}^\top$ from right and left, respectively, yields

$$A_\Lambda\tilde{U}^\top X\tilde{U}A_\Lambda^\top - \tilde{U}^\top X\tilde{U} + BB^\top = 0, \tag{C.4}$$

where we used the fact that B in (2.6) has the property that $\tilde{U}^\top BB^\top\tilde{U} = BB^\top$.

Equation (C.4) shows that $\tilde{U}^\top X\tilde{U}$ satisfies the Lyapunov equation when A_Q is replaced by A_Λ. Next, we show that after after substituting Q by Λ, the robustness $\mathcal{J}(\alpha,\beta)$ would stay the same, i.e., $H_2^2(A_\Lambda, B, \sqrt{\frac{1}{2}}\Lambda^{1/2}T) = H_2^2(A_Q, B, RT)$, where $R = \sqrt{\frac{1}{2}}\Lambda^{1/2}U^\top$. To show this, note that from (3.8), we have

$$\begin{aligned} H_2^2\left(A_\Lambda, B, \sqrt{\frac{1}{2}}\Lambda^{1/2}T\right) &= \frac{1}{2}\operatorname{Tr}((\Lambda^{1/2}T)\tilde{U}^\top X\tilde{U}(\Lambda^{1/2}T)^\top) \\ &= \frac{1}{2}\operatorname{Tr}((\Lambda^{1/2}T\tilde{U}^\top)X(\Lambda^{1/2}T\tilde{U}^\top)^\top) = \operatorname{Tr}((RT)X(RT)^\top), \end{aligned}$$

where the last equality is true as $T\tilde{U}^\top = U^\top T$ for T and U given in (3.11) and (C.2). This result completes the proof of our claim that we can assume Q is diagonal. For simplicity we will continue our analysis with A_Q, assuming its a diagonal matrix.

Let P_π be the permutation matrix associated with the permutation π over the set $\{1, 2, \ldots, 2d\}$ that satisfies $\pi(i) = 2i-1$ for $1 \le i \le d$ and $\pi(i) = 2(i-d)$ for $d+1 \le i \le 2d$. It is well-known that permutation matrices satisfy $P_\pi^{-1} = P_\pi^\top = P_{\pi^{-1}}$; therefore, multiplying Lyapunov equation (3.9) by P_π and $P_\pi^\top$ from left and right, respectively, leads to

$$(P_\pi A_Q P_\pi^\top)Y(P_\pi A_Q^\top P_\pi^\top) - Y + P_\pi BB^\top P_\pi^\top = 0, \tag{C.5}$$

where $Y = P_\pi X P_\pi^\top$. It follows from (3.28) that

$$P_\pi A_Q P_\pi^\top = \mathbf{diag}([T_i]_{i=1}^d) \text{ and } T_i = \begin{bmatrix} (1+\beta)(1-\alpha\lambda_i) & -\beta(1-\alpha\lambda_i) \\ 1 & 0 \end{bmatrix},\ i = 1, \ldots, d,$$

and $0 < \mu = \lambda_1 \le \lambda_2 \le \cdots \le \lambda_d = L$ are the eigenvalues of Q. Since $BB^T = \left[\begin{smallmatrix} \alpha^2 I_d & 0_d \\ 0_d & 0_d \end{smallmatrix}\right]$, $P_\pi BB^\top P_\pi^\top$ is a $2d$ by $2d$ diagonal matrix with α^2 on entries $(1,1), (3,3), \ldots, (2d-1, 2d-1)$ and zero elsewhere. Hence, Y that solves (C.5) is a block diagonal matrix in the form $Y = \mathbf{diag}([Y_i]_{i=1}^d)$, where $Y_i = \left[\begin{smallmatrix} y_i^u & y_i^o \\ y_i^o & y_i^d \end{smallmatrix}\right]$ satisfies the equality

$$\begin{bmatrix} (1+\beta)(1-\alpha\lambda_i) & -\beta(1-\alpha\lambda_i) \\ 1 & 0 \end{bmatrix} Y_i \begin{bmatrix} (1+\beta)(1-\alpha\lambda_i) & 1 \\ -\beta(1-\alpha\lambda_i) & 0 \end{bmatrix} - Y_i + \begin{bmatrix} \alpha^2 & 0 \\ 0 & 0 \end{bmatrix} = 0$$

for all $i = 1, \ldots, d$. This is equivalent to the linear system

$$\begin{bmatrix} (1+\beta)^2(1-\alpha\lambda_i)^2 - 1 & -2\beta(1+\beta)(1-\alpha\lambda_i)^2 & \beta^2(1-\alpha\lambda_i)^2 \\ (1+\beta)(1-\alpha\lambda_i) & -1-\beta(1-\alpha\lambda_i) & 0 \\ 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} y_i^u \\ y_i^o \\ y_i^d \end{bmatrix} = \begin{bmatrix} -\alpha^2 \\ 0 \\ 0 \end{bmatrix}.$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 05/03/20 to 18.9.61.111. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Solving this system of equations, we obtain

$$y_i^u = y_i^d = \alpha \frac{1+\beta(1-\alpha\lambda_i)}{\lambda_i(1-\beta(1-\alpha\lambda_i))(2+2\beta-\alpha\lambda_i(1+2\beta))}, \qquad y_i^o = \frac{\alpha^2(1+\beta)(1-\alpha\lambda_i)}{\alpha\lambda_i(1-\beta(1-\alpha\lambda_i))(2+2\beta-\alpha\lambda_i(1+2\beta))}. \tag{C.6}$$

The $\mathcal{J}(\alpha,\beta)$ can be computed using

$$\operatorname{Tr}\left(\left(\sqrt{\frac{1}{2}}Q^{1/2}T\right)X\left(\sqrt{\frac{1}{2}}Q^{1/2}T\right)^\top\right) = \frac{1}{2}\operatorname{Tr}\left(P_\pi T^\top Q T P_\pi^\top Y\right). \tag{C.7}$$

The matrix $P_\pi T^\top Q T P_\pi^\top$ is block diagonal with 2×2 matrices $\left[\begin{smallmatrix}\lambda_i & 0\\ 0 & 0\end{smallmatrix}\right]$ on its diagonal. Therefore, using (C.6), the robustness measure $\mathcal{J}(\alpha,\beta)$ is equal to

$$\frac{1}{2}\sum_{i=1}^{d}\alpha\frac{1+\beta(1-\alpha\lambda_i)}{(1-\beta(1-\alpha\lambda_i))(2+2\beta-\alpha\lambda_i(1+2\beta))}. \tag{C.8}$$

Appendix D. Convexity of $u_{\alpha,\beta}(\lambda)$. We next show that $u_{\alpha,\beta}(\lambda)$ appearing in the definition of the $\mathcal{J}(\alpha,\beta)$ for the AG algorithm is convex with respect to λ.

LEMMA D.1. *Let $(\alpha,\beta)\in\mathcal{S}$ where $\mathcal{S}$ is the stability region of the dynamical system representation of AG given by* (3.30). *The function $u_{\alpha,\beta}(\lambda)$ defined by* (3.35) *is convex on the interval $[\mu, L]$.*

Proof. The function $u_{\alpha,\beta}(\lambda)$ can be written in terms of $\tilde{\lambda} := \beta(1-\alpha\lambda)$ as follows:

$$q_\lambda(\tilde{\lambda}) = \frac{\alpha}{2}\frac{1+\tilde{\lambda}}{(1-\tilde{\lambda})(1+\tilde{\lambda}\gamma)}, \tag{D.1}$$

where $\gamma := 2+\frac{1}{\beta}$. It follows from (3.30) that for $(\alpha,\beta)\in\mathcal{S}$, $1\geq\tilde{\lambda}\geq\frac{-1}{\gamma}$, and thus both terms in denominator of (D.1) are positive. Note that $\tilde{\lambda}$ is linear, and since the composition of a convex functions with a linear function is convex, it suffices to show $q_\lambda(\tilde{\lambda})$ is a convex function of $\tilde{\lambda}$ over domain $[\frac{-1}{\gamma},1]$. To show this, we simply compute the second derivative of (D.1) with respect to $\tilde{\lambda}$. After doing some algebra,

$$\frac{d}{d\tilde{\lambda}^2}q_\lambda(\tilde{\lambda}) = \frac{\alpha}{2}\left(\frac{2(\gamma^3-\gamma^2)}{(\gamma+1)(\gamma\tilde{\lambda}+1)^3}+\frac{4}{(\gamma+1)(1-\tilde{\lambda})^3}\right), \tag{D.2}$$

which is nonnegative as $\gamma\geq 2$ and $\tilde{\lambda}\in[\frac{-1}{\gamma},1]$. This completes the proof. ∎

Appendix E. Defining rate and robustness based on iterates. In this supplementary file, we first recall how an alternative robustness measure can be defined based on the distance of the iterates to the optimal solution instead of the robustness measure $\mathcal{J}$ we introduced in the main text based on the asymptotic expected suboptimality in function values. Here, we focus on the iterate sequence $\{x_k\}_k$ to characterize the notions of rate and robustness. First we consider the case that f is a quadratic function in the form of $f(x)=\frac{1}{2}x^\top Qx-p^\top x+r$. Using (3.4) along with the relation $x_k = T\xi_k$ with T defined in section 3.1, for both GD and AG the sequence $\{\mathbb{E}[x_k]\}_k$ goes to zero with rate $\rho(A_Q)$.

However, due to the noise injected at each step, the limit of the sequence $\{x_k\}$ will oscillate around the optimal solution with a nonzero variance. Thus, a natural

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 05/03/20 to 18.9.61.111. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

metric to measure robustness is then to study the *asymptotic normalized variance* by considering the limit $\mathcal{J}' = \lim_{k\to\infty} \frac{1}{\sigma^2}\mathbb{E}[\|x_k - x^*\|^2]$. A similar line of argument as Lemma 3.1 shows that this limit exists and is in fact equal to $H_2^2(A_Q, B, T)$. This quantity can be viewed as the *robustness to noise in terms of iterates* because it is equal to the ratio of the power of the iterates to the power of the input noise, measuring how much a system amplifies input noise. In particular, the smaller this measure is, the more robust the system is under additive random noise.

The robustness $\mathcal{J}'$ can be evaluated precisely for GD and AG methods just as we did in section 3 for $\mathcal{J}$. For the GD method with constant stepsize $\alpha \in (0, 2/L)$, the robustness to noise in terms of iterates is denoted as $\mathcal{J}'(\alpha)$ to show the dependence to α. The following proposition, which can be proved similarly to Proposition 3.2, shows the explicit characterization of $\mathcal{J}'(\alpha)$.

PROPOSITION E.1. *Let f be a quadratic function of the form $f(x) = \frac{1}{2}x^\top Qx - p^\top x + r$. Consider the GD iterations given by (2.3) with constant stepsize $\alpha \in (0, 2/L)$. Then the robustness of the GD method in terms of iterates is given by*

$$\mathcal{J}'(\alpha) = \alpha^2 \sum_{i=1}^{d} \frac{1}{1-(1-\alpha\lambda_i)^2} = \alpha \sum_{i=1}^{d} \frac{1}{\lambda_i(2-\alpha\lambda_i)}, \tag{E.1}$$

where $0 < \mu = \lambda_1 \le \lambda_2 \le \dots \lambda_d = L$ are the eigenvalues of Q.

For AG, with constant stepsize α and momentum parameter β, we denote the robustness to noise in terms of iterates as $\mathcal{J}'(\alpha, \beta)$. The following theorem, which can be proved similarly to Proposition 3.5, provides an explicit formula for $\mathcal{J}(\alpha, \beta)$ in terms of the eigenvalues of Q.

PROPOSITION E.2. *Let f be a quadratic function of the form $f(x) = \frac{1}{2}x^\top Qx - p^\top x + r$. Consider the AG iterations given by (2.5) with parameters $(\alpha, \beta) \in \mathcal{S}$. Then the robustness of the AG method in terms of iterates is given by*

$$\mathcal{J}'(\alpha, \beta) = \sum_{i=1}^{d} u'_{\alpha,\beta}(\lambda_i), \tag{E.2}$$

where $\mu = \lambda_1 \le \lambda_2 \le \cdots \le \lambda_d = L$ are the eigenvalues of Q and

$$u'_{\alpha,\beta}(\lambda) \triangleq \alpha \frac{1+\beta(1-\alpha\lambda)}{\lambda(1-\beta(1-\alpha\lambda))(2+2\beta-\alpha\lambda(1+2\beta))}. \tag{E.3}$$

As discussed in section 3, the $\mathcal{J}'(\alpha, \beta)$ admits a tractable upper bound in the form of $\mathcal{J}'(\alpha, \beta) \le d \max(u'_{\alpha,\beta}(\mu), u'_{\alpha,\beta}(L))$ which only depends on μ and L.

We can also extend the definitions of rate and robustness in terms of iterates to the case that $f \in S_{\mu,L}(\mathbb{R}^d)$. Note that in general the sequence $\{x_k\}_k$ might not converge to the optimal solution in expectation (see [14]). Using the family of Lyapunov functions $V_{P,c}(\xi)$, it can be shown that, similar to (4.3), the following inequality holds for both GD and AG with properly chose parameters:

$$\mathbb{E}[\|x_k - x^*\|^2] \le \rho^{2k}\psi'_0 + \sigma^2 R', \quad k \ge 1, \tag{E.4}$$

where $0 < \rho < 1$ is the same ρ as (4.3) and also ψ'_0 and R' are nonnegative numbers and depend on algorithm parameters and initial point x_0. For instance, Proposition 4.3 implies that (E.4) holds for GD, i.e., for all $k \ge 0$,

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

$$\mathbb{E}[\|x_k - x^*\|^2] \leq \rho(\alpha)^{2k} \|x_0 - x^*\|^2 + \sigma^2 R'(\alpha), \quad \text{where} \tag{E.5}$$

$$R'(\alpha) \triangleq \frac{\alpha^2 d}{1 - \rho(\alpha)^2}. \tag{E.6}$$

Similarly, we can derive (E.4) for AG by using Proposition 4.6.

Finally, we show that the $R'(\alpha)$ is a tight bound for $\mathcal{J}'$. For quadratic $f \in S_{\mu,L}(\mathbb{R}^d)$, $\mathcal{J}'$ can be written in closed form as in (E.1). Note that if $\lambda_i = \mu = L$ for $i = 1, \ldots, d$, then this quantity is equal to $R'(\alpha)$.

REFERENCES

[1] N. S. Aybat, A. Fallah, M. Gürbüzbalaban, and A. Ozdaglar, *A Universally Optimal Multistage Accelerated Stochastic Gradient Method*, https://arxiv.org/abs/1901.08022, 2019.

[2] F. Bach and E. Moulines, *Non-strongly-convex smooth stochastic approximation with convergence rate o(1/n)*, in Advances in Neural Information Processing Systems 26, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, eds., Curran Associates, 2013, pp. 773–781.

[3] R. Bassily, A. Smith, and A. Thakurta, *Private empirical risk minimization: Efficient algorithms and tight error bounds*, in Proceedings of the 55th Annual Symposium on Foundations of Computer Science, IEEE, 2014, pp. 464–473.

[4] D. Bertsekas, *Nonlinear Programming*, Athena Scientific, Belmont, MA, 1999.

[5] D. Bertsekas, *Incremental gradient, subgradient, and proximal methods for convex optimization: A survey*, Optim. Machine Learning, 2010 (2011), p. 3.

[6] B. Can, M. Gurbuzbalaban, and L. Zhu, *Accelerated linear convergence of stochastic momentum methods in Wasserstein distances*, in Proceedings of the 36th International Conference on Machine Learning, K. Chaudhuri and R. Salakhutdinov, eds., in Proceedings of Machine Learning Research, Long Beach, CA, PMLR 97, 2019, pp. 891–901, http://proceedings.mlr.press/v97/can19a.html.

[7] B. Can, M. Gürbüzbalaban, and L. Zhu, *Accelerated Linear Convergence of Stochastic Momentum Methods in Wasserstein Distances*, https://arxiv.org/abs/1901.07445, 2019.

[8] X. Cheng, N. S. Chatterji, P. L. Bartlett, and M. I. Jordan, *Underdamped Langevin MCMC: A Non-Asymptotic Analysis*, https://arxiv.org/abs/1707.03663, 2017.

[9] S. Cyrus, B. Hu, B. Van Scoy, and L. Lessard, *A robust accelerated optimization algorithm for strongly convex functions*, in Proceedings of the Annual American Control Conference, 2018, pp. 1376–1381, https://doi.org/10.23919/ACC.2018.8430824.

[10] A. d'Aspremont, *Smooth optimization with approximate gradient*, SIAM J. Optim., 19 (2008), pp. 1171–1183, https://doi.org/10.1137/060676386.

[11] O. Devolder, *Exactness, Inexactness and Stochasticity in First-Order Methods for Large-Scale Convex Optimization*, Ph.D. thesis, ICTEAM and CORE, Université Catholique de Louvain, 2013.

[12] O. Devolder, F. Glineur, and Y. Nesterov, *Intermediate Gradient Methods for Smooth Convex Problems with Inexact Oracle*, Tech. report, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), 2013.

[13] O. Devolder, F. Glineur, and Y. Nesterov, *First-order methods of smooth convex optimization with inexact oracle*, Math. Program., 146 (2014), pp. 37–75.

[14] A. Dieuleveut, A. Durmus, and F. Bach, *Bridging the Gap Between Constant Step Size Stochastic Gradient Descent and Markov Chains*, https://arxiv.org/abs/1707.06386, 2017.

[15] P. Dvurechensky and A. Gasnikov, *Stochastic intermediate gradient method for convex problems with stochastic inexact oracle*, J. Optim. Theory Appl., 171 (2016), pp. 121–145.

[16] A. Eberle, A. Guillin, and R. Zimmer, *Couplings and Quantitative Contraction Rates for Langevin Dynamics*, https://arxiv.org/abs/1703.01617, 2017.

[17] A. Edelman and H. Murakami, *Polynomial roots from companion matrix eigenvalues*, Math. Comp., 64 (1995), pp. 763–776.

[18] M. Fazlyab, A. Ribeiro, M. Morari, and V. Preciado, *Analysis of optimization algorithms via integral quadratic constraints: Nonstrongly convex problems*, SIAM J. Optim., 28 (2018), pp. 2654–2689, https://doi.org/10.1137/17M1136845.

[19] N. Flammarion and F. Bach, *From averaging to acceleration, there is only a step-size*, in Proceedings of the Conference on Learning Theory, 2015, pp. 658–695.

[20] W. H. Fleming and M. James, *The risk-sensitive index and the H_2 and H_∞, norms for nonlinear systems*, Math. Control Signals Systems, 8 (1995), pp. 199–221.

Downloaded 05/03/20 to 18.9.61.111. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

[21] X. Gao, M. Gürbüzbalaban, and L. Zhu, *Breaking Reversibility Accelerates Langevin Dynamics for Global Non-Convex Optimization*, https://arxiv.org/abs/1812.07725, 2018.
[22] X. Gao, M. Gürbüzbalaban, and L. Zhu, *Global Convergence of Stochastic Gradient Hamiltonian Monte Carlo for Non-Convex Stochastic Optimization: Non-Asymptotic Performance Bounds and Momentum-Based Acceleration*, https://arxiv.org/abs/1809.04618, 2018.
[23] S. Ghadimi and G. Lan, *Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization* I*: A generic algorithmic framework*, SIAM J. Optim., 22 (2012), pp. 1469–1492, https://doi.org/10.1137/110848864.
[24] S. Ghadimi and G. Lan, *Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization* II*: Shrinking procedures and optimal algorithms*, SIAM J. Optim., 23 (2013), pp. 2061–2089, https://doi.org/10.1137/110848876.
[25] M. Grant and S. Boyd, *CVX: MATLAB Software for Disciplined Convex Programming, Version* 2.1, http://cvxr.com/cvx, 2014.
[26] W. Haddad and D. Bernstein, *Parameter-dependent Lyapunov functions and the discrete-time Popov criterion for robust analysis*, Automatica, 30 (1994), pp. 1015–1021.
[27] M. Hardt, *Robustness Versus Acceleration*, http://blog.mrtz.org/2014/08/18/robustness-versus-acceleration, 2014.
[28] B. Hu and L. Lessard, *Dissipativity theory for Nesterov's accelerated method*, in Proceedings of the 34th International Conference on Machine Learning, PMLR 70, 2017, pp. 1549–1557.
[29] B. Hu, P. Seiler, and L. Lessard, *Analysis of Approximate Stochastic Gradient Using Quadratic Constraints and Sequential Semidefinite Programs*, https://arxiv.org/abs/1711.00987, 2017.
[30] A. Jadbabaie and A. Olshevsky, *Combinatorial bounds and scaling laws for noise amplification in networks*, in Proceedings of the European Control Conference, IEEE, 2013, pp. 596–601.
[31] A. Jadbabaie and A. Olshevsky, *On performance of consensus protocols subject to noise: Role of hitting times and network structure*, in Proceedings of the 55th Conference on Decision and Control, IEEE, 2016, pp. 179–184.
[32] G. Lan, *An optimal method for stochastic composite optimization*, Math. Program., 133 (2012), pp. 365–397, https://doi.org/10.1007/s10107-010-0434-y.
[33] L. Lessard, B. Recht, and A. Packard, *Analysis and design of optimization algorithms via integral quadratic constraints*, SIAM J. Optim., 26 (2016), pp. 57–95.
[34] S. Michalowsky and C. Ebenbauer, *The multidimensional n-th order heavy ball method and its application to extremum seeking*, in Proceedings of the 53rd Conference on Decision and Control, IEEE, 2014, pp. 2660–2666.
[35] H. Mohammadi, M. Razaviyayn, and M. R. Jovanović, *Variance amplification of accelerated first-order algorithms for strongly convex quadratic optimization problems*, in Proceedings of the IEEE Conference on Decision and Control, IEEE, 2018, pp. 5753–5758.
[36] Y. Nesterov, *A method of solving a convex programming problem with convergence rate* $o(1/k^2)$, Soviet Math. Dokl., 27 (1983), pp. 372–376.
[37] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Appl. Optim. 87, Springer, New York, 2004, https://doi.org/10.1007/978-1-4419-8853-9.
[38] B. O'Donoghue and E. Candès, *Adaptive restart for accelerated gradient schemes*, Found. Comput. Math., 15 (2015), pp. 715–732.
[39] B. T. Polyak and A. B. Juditsky, *Acceleration of stochastic approximation by averaging*, SIAM J. Control Optim., 30 (1992), pp. 838–855.
[40] M. Raginsky, A. Rakhlin, and M. Telgarsky, *Non-convex learning via stochastic gradient langevin dynamics: A nonasymptotic analysis*, in Proceedings of the 2017 Conference on Learning Theory, Amsterdam, 2017, PMLR 65, 2017, pp. 1674–1703.
[41] H. Robbins and S. Monro, *A stochastic approximation method*, Ann. Math. Statist., 22 (1951), pp. 400–407.
[42] S. Safavi, B. Joshi, G. França, and J. Bento, *An explicit convergence rate for Nesterov's method from sdp*, in Proceedings of the International Symposium on Information Theory, IEEE, 2018, pp. 1560–1564.
[43] M. Schmidt, N. Le Roux, and F. Bach, *Convergence rates of inexact proximal-gradient methods for convex optimization*, in Advances in Neural Information Processing Systems 24, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, eds., Curran Associates, 2011, pp. 1458–1466.
[44] A. A. Stoorvogel, *The robust* H_2 *control problem: A worst-case design*, IEEE Trans. Automat. Control, 38 (1993), pp. 1358–1371, https://doi.org/10.1109/9.237647.

Downloaded 05/03/20 to 18.9.61.111. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

[45] A. Wilson, B. Recht, and M. Jordan, *A Lyapunov Analysis of Momentum Methods in Optimization*, https://arxiv.org/abs/1611.02635, 2016.
[46] K. Zhou, J. C. Doyle, and K. Glover, *Robust and Optimal Control*, Prentice Hall, Englewood Cliffs, NJ, 1996.

Downloaded 05/03/20 to 18.9.61.111. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.