Mathematical Programming
https://doi.org/10.1007/s10107-020-01486-1

FULL LENGTH PAPER

Series A

Analysis of biased stochastic gradient descent using sequential semidefinite programs

Bin Hu[1] · Peter Seiler[2] · Laurent Lessard[3]

Received: 25 June 2018 / Accepted: 2 March 2020

Abstract

We present a convergence rate analysis for biased stochastic gradient descent (SGD), where individual gradient updates are corrupted by computation errors. We develop stochastic quadratic constraints to formulate a small linear matrix inequality (LMI) whose feasible points lead to convergence bounds of biased SGD. Based on this LMI condition, we develop a sequential minimization approach to analyze the intricate trade-offs that couple stepsize selection, convergence rate, optimization accuracy, and robustness to gradient inaccuracy. We also provide feasible points for this LMI and obtain theoretical formulas that quantify the convergence properties of biased SGD under various assumptions on the loss functions.

Keywords Biased stochastic gradient · Robustness to inexact gradient · Convergence rates · Convex optimization · First-order methods

This work is supported by the NSF Awards 1656951, 1750162, 1254129, and the NASA Langley NRA Cooperative Agreement NNX12AM55A. Bin Hu and Laurent Lessard also acknowledge support from the Wisconsin Institute for Discovery, the College of Engineering, and the Department of Electrical and Computer Engineering at the University of Wisconsin–Madison.

✉ Bin Hu
binhu7@illinois.edu
https://binhu7.github.io/

Peter Seiler
pseiler@umich.edu

Laurent Lessard
laurent.lessard@wisc.edu
https://laurentlessard.com

[1] Department of Electrical and Computer Engineering, University of Illinois at Urbana–Champaign, 306 N Wright St., Urbana, IL 61801, USA

[2] Department of Electrical Engineering and Computer Science, University of Michigan, 1301 Beal Avenue, Ann Arbor, MI 48109, USA

[3] Department of Electrical and Computer Engineering, Wisconsin Institute for Discovery, University of Wisconsin–Madison, 330 N. Orchard St., Madison, WI 53715, USA

Published online: 23 March 2020

Mathematics Subject Classification 90C06 · 90C15 · 90C25 · 90C30 · 68Q25 · 62L20

1 Introduction

Empirical risk minimization (ERM) is a prevalent topic in machine learning research [7,36]. Ridge regression, ℓ_2-regularized logistic regression, and support vector machines (SVM) can all be formulated as the following ERM problem

$$\min_{x\in\mathbb{R}^p} g(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x), \tag{1}$$

where $g : \mathbb{R}^p \to \mathbb{R}$ is the objective function. Stochastic gradient descent (SGD) [4,6,26] has been widely used for ERM to exploit redundancy in the training data. The SGD method applies the update rule

$$x_{k+1} = x_k - \alpha_k u_k, \tag{2}$$

where $u_k = \nabla f_{i_k}(x_k)$ and the index i_k is uniformly sampled from $\{1, 2, \ldots, n\}$ in an independent and identically distributed (IID) manner. The convergence properties of SGD are well understood. Under strong convexity of g and smoothness of f_i, SGD with a diminishing stepsize converges sublinearly, while SGD with a constant stepsize converges linearly to a ball around the optimal solution [15,22–24]. In the latter case, epochs can be used to balance convergence rate and optimization accuracy. Some recently-developed stochastic methods such as SAG [27,28], SAGA [11], Finito [12], SDCA [30], and SVRG [19] converge linearly with low iteration cost when applied to (1), though SGD is still popular because of its simple iteration form, low memory footprint, and nice generalization property. SGD is also commonly used as an initialization for other algorithms [27,28].

In this paper, we present a general analysis for biased SGD. This is a version of SGD where the gradient updates $\nabla f_{i_k}(x_k)$ are corrupted by additive as well as multiplicative noise. In practice, such errors can be introduced by sources such as: inaccurate numerical solvers, digital round-off errors, quantization, or sparsification. The biased SGD update equation is given by

$$x_{k+1} = x_k - \alpha_k(u_k + e_k). \tag{3}$$

Here, $u_k = \nabla f_{i_k}(x_k)$ is the individual gradient update and e_k is an error term. We consider the following error model, which unifies the error models in [3]:

$$\|e_k\|^2 \le \delta^2\|u_k\|^2 + c^2, \tag{4}$$

where $\delta \ge 0$ and $c \ge 0$ bound the relative error and the absolute error in the oracle computation, respectively. If $\delta = c = 0$, then $e_k = 0$ and we recover the standard SGD setup. The model (4) unifies the error models in [3] since:

1. If $c = 0$, then (4) reduces to a relative error model, i.e.

$$\|e_k\| \leq \delta \|u_k\| \tag{5}$$

2. If $\delta = 0$, then e_k is a bounded absolute error, i.e.

$$\|e_k\| \leq c \tag{6}$$

We assume that both δ and c are known in advance. We make no assumptions about how e_k is generated, just that it satisfies (4). Thus, we will seek a worst-case bound that holds regardless of whether e_k is random, set in advance, or chosen adversarially.

Suppose the cost function g admits a unique minimizer $x_\star$. For standard SGD (without computation error), u_k is an unbiased estimator of $\nabla g(x_k)$. Hence under many circumstances, one can control the final optimization error $\|x_k - x_\star\|$ by decreasing the stepsize α_k. Specifically, suppose g is m-strongly convex. Under various assumptions on f_i, one can prove the following typical bound for standard SGD with a constant stepsize α [5,22,24]:

$$\mathbb{E}\|x_k - x_\star\|^2 \leq \rho^{2k}\mathbb{E}\|x_0 - x_\star\|^2 + H_\star \tag{7}$$

where $\rho^2 = 1 - 2m\alpha + O(\alpha^2)$ and $H_\star = O(\alpha)$. By decreasing stepsize α, one can control the final optimization error $H_\star$ at the price of slowing down the convergence rate ρ. The convergence behavior of biased SGD is different. Since the error term e_k can be chosen adversarially, the sum $(u_k + e_k)$ may no longer be an unbiased estimator of $\nabla g(x_k)$. The error term e_k may introduce a bias which cannot be overcome by decreasing stepsize α. Hence the final optimization error in biased SGD heavily depends on the error model of e_k. In this paper, we quantify the convergence properties of biased SGD (3) with the error model (4) using worst-case analysis.

Main contribution. The main novelty of this paper is that our analysis simultaneously addresses the relative error and the absolute error in the gradient computation. We formulate a linear matrix inequality (LMI) that directly leads to convergence bounds of biased SGD and couples the relationship between δ, c, α_k and the assumptions on f_i. This convex program can be solved both numerically and analytically to obtain various convergence bounds for biased SGD. Based on this LMI, we develop a sequential minimization approach that can analyze biased SGD with an arbitrary time-varying stepsize. We also obtain analytical rate bounds in the form of (7) for biased SGD with constant stepsize. However, our bound requires $\rho^2 = 1 - \frac{m^2-\delta^2\tilde{M}}{m}\alpha + O(\alpha^2)$ [1] and $H_\star = \frac{c^2+2\delta^2G^2}{m^2-\delta^2\tilde{M}} + O(\alpha)$ where $\tilde{M}$ and G^2 are some prescribed constants determined by the assumptions on f_i. Based on this result, there is no way to shrink $H_\star$ to 0. This is consistent with our intuition since the gradient estimator as well as the final optimization result can be biased. We show that this "uncontrollable" biased optimization

[1] When $\delta = c = 0$, this rate bound does not reduce to $\rho^2 = 1 - 2m\alpha + O(\alpha^2)$. This is due to the inherent differences between the analyses of biased SGD and the standard SGD. See Remark 4 for a detailed explanation.

error is $\frac{c^2+2\delta^2G^2}{m^2-\delta^2\tilde{M}}$. The resultant analytical rate bounds highlight the design trade-offs for biased SGD.

The work in this paper complements the ongoing research on stochastic optimization methods, which mainly focuses on the case where the oracle computation is exact. The stepsize selection in biased SGD must address the trade-offs between speed, accuracy, and inexactness in the oracle computations. Our analysis brings new theoretical insights for understanding such trade-offs in the presence of biased gradient computation. It is also worth mentioning that the robustness of full gradient methods with respect to gradient inexactness has been extensively studied [9,13,29]. However, addressing a unified error model that combines the absolute error and the relative error is still non-trivial. Our analysis complements the existing results in [9,13,29] by providing a unified treatment of the error model (4). Notice that it is important to include the relative error model in the analysis since it covers the numerical round-off error as a special case. If one treats the round-off error as an absolute error with time-varying c_k, then the specific value of c_k will depend on the state x_k and can not be fixed beforehand. In contrast, if one models the round-off error as a relative error, the value δ can be fixed as a constant beforehand.

The approach taken in this paper can be viewed as a stochastic extension of the work in [21,25] that analyzes the linear convergence rates of deterministic optimization methods (gradient descent, Nesterov's method, ADMM, etc.) using quadratic constraints and semidefinite programs. Notice that the analysis for (deterministic) biased gradient descent in [21] is numerical. In this paper, we derive analytical formulas quantifying the convergence properties of the biased SGD. It is worth mentioning that one can combine jump system theory with quadratic constraints to analyze SAGA, Finito, and SDCA in a unified manner [18]. However, the analysis in [18] does not directly address the trade-offs between the convergence speed ρ^2 and the optimization error $H_\star$, and cannot be easily tailored for biased SGD. Another related line of work that uses semidefinite programs to analyze optimization methods is built upon the idea of formulating worst-case analysis as the so-called performance estimation problem (PEP) [14,33,34]. It is recognized that there is a fundamental connection between the quadratic constraint approach and the PEP framework [35]. Recently, the PEP framework in [14,33,34] has been extended for the stochastic setup [32]. In addition, it is known that the PEP approach can also be applied to study the bias in the (deterministic) gradient descent method [10]. It is possible to extend the results in [10,32] for a PEP-based analysis of biased SGD. This is an interesting topic for future research.

The rest of the paper is organized as follows. In Sect. 2, we formulate LMI testing conditions for convergence analysis of biased SGD. The resultant LMIs are then solved sequentially, yielding recursive convergence bounds for biased SGD. In Sect. 3, we simplify the analytical solutions of the resultant sequential LMIs and derive analytical rate bounds in the form of (7) for biased SGD with a constant stepsize. Our results highlight various design trade-offs for biased SGD. Finally, we show how existing results on standard SGD (without gradient computation error) can be recovered using our proposed LMI approach, and discuss how a time-varying stepsize can potentially impact the convergence behaviors of biased SGD (Sect. 4).

1.1 Notation

The $p \times p$ identity matrix and the $p \times p$ zero matrix are denoted as I_p and 0_p, respectively. The subscript p is occasionally omitted when the dimensions are clear by the context. When a matrix P is negative semidefinite, we will use the notation $P \preceq 0$. The Kronecker product of two matrices A and B is denoted $A \otimes B$.

Definition 1 (*Smooth functions*) A differentiable function $f : \mathbb{R}^p \to \mathbb{R}$ is L-smooth for some $L > 0$ if the following inequality is satisfied:

$$\|\nabla f(x) - \nabla f(y)\| \le L\,\|x - y\| \qquad \text{for all } x, y \in \mathbb{R}^p.$$

Definition 2 (*Convex functions*) Let $\mathcal{F}(m, L)$ for $0 \le m \le L \le \infty$ denote the set of differentiable functions $f : \mathbb{R}^p \to \mathbb{R}$ satisfying the following inequality for all $x, y \in \mathbb{R}^p$.

$$\begin{bmatrix} x - y \\ \nabla f(x) - \nabla f(y) \end{bmatrix}^\mathsf{T} \begin{bmatrix} -2mI_p & (1+\frac{m}{L})I_p \\ (1+\frac{m}{L})I_p & -\frac{2}{L}I_p \end{bmatrix} \begin{bmatrix} x - y \\ \nabla f(x) - \nabla f(y) \end{bmatrix} \ge 0. \tag{8}$$

Note that $\mathcal{F}(0, \infty)$ is the set of all convex functions, $\mathcal{F}(0, L)$ is the set of all convex L-smooth functions, $\mathcal{F}(m, \infty)$ with $m > 0$ is the set of all m-strongly convex functions, and $\mathcal{F}(m, L)$ with $m > 0$ is the set of all m-strongly convex and L-smooth functions. If $f \in \mathcal{F}(m, L)$ with $m > 0$, then f has a unique global minimizer.

Definition 3 Let $\mathcal{S}(m, L)$ for $0 \le m \le L \le \infty$ denote the set of differentiable functions $g : \mathbb{R}^p \to \mathbb{R}$ having some global minimizer $x_\star \in \mathbb{R}^p$ and satisfying the following inequality for all $x, y \in \mathbb{R}^p$.

$$\begin{bmatrix} x - x_\star \\ \nabla g(x) \end{bmatrix}^\mathsf{T} \begin{bmatrix} -2mI_p & (1+\frac{m}{L})I_p \\ (1+\frac{m}{L})I_p & -\frac{2}{L}I_p \end{bmatrix} \begin{bmatrix} x - x_\star \\ \nabla g(x) \end{bmatrix} \ge 0. \tag{9}$$

If $g \in \mathcal{S}(m, L)$ with $m > 0$, then $x_\star$ is also the unique stationary point of g. It is worth noting that $\mathcal{F}(m, L) \subset \mathcal{S}(m, L)$. In general, a function $g \in \mathcal{S}(m, L)$ may not be convex. If $g \in \mathcal{S}(m, \infty)$, then g may not be smooth. The condition (9) is similar to the notion of *one-point convexity* [2,8,31] and *star-convexity* [20].

1.2 Assumptions

Referring to the problem setup (1), we will adopt the general assumption that $g \in \mathcal{S}(m, \infty)$ with $m > 0$. So in general, g may not be convex. We will analyze four different cases, characterized by different assumptions on individual f_i: (I) Bounded shifted gradients:[2] $\|\nabla f_i(x) - mx\| \le \beta$ for all $x \in \mathbb{R}^p$; (II) f_i is L-smooth; (III) $f_i \in \mathcal{F}(0, L)$; (IV) $f_i \in \mathcal{F}(m, L)$.

[2] This case is a variant of the common assumption $\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x)\|^2 \le \beta$. One can check that this case holds for several ℓ_2-regularized problems including SVM and logistic regression.

Assumption I is a natural assumption for SVM[3] and logistic regression while Assumptions II, III, or IV can be used for ridge regression, logistic regression, and smooth SVM. The m assumed in cases I and IV is the same as the m used in the assumption on $g \in \mathcal{S}(m, \infty)$.

2 Analysis framework

2.1 An LMI condition for the analysis of biased SGD

To analyze the convergence properties of biased SGD, we present a small linear matrix inequality (LMI) whose feasible points directly lead to convergence bounds of the biased SGD (3) with the error model (4).

Theorem 1 (Main Theorem) *Consider biased SGD* (3) *with* $g \in \mathcal{S}(m, \infty)$ *for some* $m > 0$*, and let* $x_\star$ *be the unique global minimizer of* g*. Given one of the four conditions on* f_i *and the corresponding* $M = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix}$ *and* G *from Table* 1*, if the following holds for some choice of nonnegative* $\lambda_k, \nu_k, \mu_k, \rho_k$*,*

$$\begin{bmatrix} -\rho_k^2 - 2\nu_k m + \lambda_k M_{11} & \nu_k + \lambda_k M_{12} & -1 & 0 \\ \nu_k + \lambda_k M_{21} & \mu_k \delta^2 + \lambda_k M_{22} & \alpha_k & 0 \\ -1 & \alpha_k & -1 & \alpha_k \\ 0 & 0 & \alpha_k & -\mu_k \end{bmatrix} \preceq 0 \tag{10}$$

where the inequality is taken in the semidefinite *sense, then the biased SGD iterates satisfy*

$$\mathbb{E}\|x_{k+1} - x_\star\|^2 \le \rho_k^2\, \mathbb{E}\|x_k - x_\star\|^2 + (2\lambda_k G^2 + \mu_k c^2) \tag{11}$$

Proof The proof is based on extending the quadratic constraint approach in [21] to the stochastic case. Specifically, one can show that for each of the four conditions on f_i and the corresponding M and G in Table 1, the following quadratic constraint holds.

$$\mathbb{E}\begin{bmatrix} x_k - x_\star \\ u_k \end{bmatrix}^\mathsf{T} (M \otimes I_p) \begin{bmatrix} x_k - x_\star \\ u_k \end{bmatrix} \ge -2G^2. \tag{12}$$

Then one can use some standard arguments from the controls literature to prove the statement in this theorem. A detailed proof is presented in the appendix. □

Remark 1 Under mild technical assumptions, the result in Theorem 1 can be extended for the problem in the more general form of $\min_x\{\mathbb{E} f_i(x)\}$, since its proof does not depend on the cardinality of the index set that i is sampled from. For simplicity, our paper focuses on the finite sum setup.

[3] The loss functions for SVM are non-smooth, and u_k is actually updated using the subgradient information. For simplicity, we abuse our notation and use ∇f_i to denote the subgradient of f_i for SVM problems.

Table 1 Given that $g \in \mathcal{S}(m, \infty)$, this table shows different possible assumptions about the f_i and their corresponding values of M and G^2 that will be used for our analysis

Case	Desired assumption on the f_i	Value of M	Value of G^2
I	$(f_i(x) - \frac{m}{2}\|x\|^2)$ have bounded gradients; $\|\nabla f_i(x) - mx\| \le \beta$.	$\begin{bmatrix} -m^2 & m \\ m & -1 \end{bmatrix}$	$\beta^2 + m^2 \|x_\star\|^2$
II	The f_i are L-smooth, but are not necessarily convex.	$\begin{bmatrix} 2L^2 & 0 \\ 0 & -1 \end{bmatrix}$	$\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x_\star)\|^2$
III	The f_i are convex and L-smooth; $f_i \in \mathcal{F}(0, L)$.	$\begin{bmatrix} 0 & L \\ L & -1 \end{bmatrix}$	$\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x_\star)\|^2$
IV	The f_i are m-strongly convex and L-smooth; $f_i \in \mathcal{F}(m, L)$	$\begin{bmatrix} -2mL & L+m \\ L+m & -1 \end{bmatrix}$	$\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x_\star)\|^2$

Notice (11) can be used to prove various types of convergence results. We will briefly discuss this in Remark 2 and provide more details in later sections. For a fixed δ, the matrix in (10) is linear in $(\rho_k^2, \nu_k, \mu_k, \lambda_k, \alpha_k)$, so (10) is a *linear matrix inequality* (LMI) whose feasible set is convex and can be efficiently searched using standard semidefinite program solvers. For example, one can implement the LMIs using CVX, a package for specifying and solving convex programs [16,17]. Since the matrix in (10) is even linear in α_k, so the LMI (10) can be used to study the impacts of adaptive stepsize rules on the performance of biased SGD from a theoretical viewpoint. One may also obtain analytical formulas for certain feasibility points of the LMI (10) due to its simple form. Our analytical bounds for biased SGD are based on the following result.

Corollary 1 *Choose one of the four conditions on f_i and the corresponding $M = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix}$ and G from Table 1. Also define $\tilde{M} = M_{11} + 2mM_{12}$. Consider biased SGD (3) with $g \in \mathcal{S}(m, \infty)$ for some $m > 0$, and let $x_\star$ be the unique global minimizer of g. Suppose the stepsize satisfies the bound $0 < M_{21}\alpha_k \leq 1$ [4](which is equivalent to the following upper bound on α_k for the four cases being considered in this paper).*

Case	I	II	III	IV
M_{21}	m	0	L	$L+m$
$\tilde{M} = M_{11} + 2mM_{12}$	m^2	$2L^2$	$2mL$	$2m^2$
α_k bound	$\frac{1}{m}$	∞	$\frac{1}{L}$	$\frac{1}{L+m}$

Then biased SGD (3) with the error model (4) satisfies the bound (11) with the following nonnegative parameters

$$\mu_k = \alpha_k^2(1 + \zeta_k^{-1}) \tag{13a}$$

$$\lambda_k = \alpha_k^2(1 + \zeta_k)(1 + \delta^2\zeta_k^{-1}) \tag{13b}$$

$$\rho_k^2 = (1 + \zeta_k)(1 - 2m\alpha_k + \tilde{M}\alpha_k^2(1 + \delta^2\zeta_k^{-1})) \tag{13c}$$

where ζ_k is a parameter that satisfies $\zeta_k > 0$ and $\zeta_k \geq \frac{\alpha_k M_{21}\delta^2}{1-\alpha_k M_{21}}$. Each choice of ζ_k yields a different bound in (11).

Proof We further define

$$\nu_k = \alpha_k(1 + \zeta_k)(1 - \alpha_k M_{21}(1 + \delta^2\zeta_k^{-1})) \tag{14}$$

We will show that (13) and (14) are a feasible solution for (10). We begin with (10) and take the Schur complement with respect to the (3, 3) entry of the matrix, leading to

[4] Ensuring such a condition in practice can be challenging for many cases since it heavily relies on the estimations of problem parameters.

$$\begin{bmatrix} 1-\rho_k^2-2\nu_k m+\lambda_k M_{11} & \nu_k+\lambda_k M_{12}-\alpha_k & -\alpha_k \\ \nu_k+\lambda_k M_{21}-\alpha_k & \mu_k\delta^2+\lambda_k M_{22}+\alpha_k^2 & \alpha_k^2 \\ -\alpha_k & \alpha_k^2 & \alpha_k^2-\mu_k \end{bmatrix} \preceq 0 \tag{15}$$

Examining the (3, 3) entry, we deduce that $\mu_k > \alpha_k^2$, for if we had equality instead, the rest of the third row and column would be zero, forcing $\alpha_k = 0$. Substituting $\mu_k = \alpha_k^2(1+\zeta_k^{-1})$ for some $\zeta_k > 0$ and taking the Schur complement with respect to the (3, 3) entry, we see (15) is equivalent to

$$\begin{bmatrix} 1-\rho_k^2-2\nu_k m+\lambda_k M_{11}+\zeta_k & \nu_k+\lambda_k M_{12}-\alpha_k(1+\zeta_k) \\ \nu_k+\lambda_k M_{21}-\alpha_k(1+\zeta_k) & \lambda_k M_{22}+\alpha_k^2(1+\zeta_k)(1+\delta^2\zeta_k^{-1}) \end{bmatrix} \preceq 0 \tag{16}$$

In (16), $\zeta_k > 0$ is a parameter that we are free to choose, and each choice yields a different set of feasible tuples $(\rho_k^2, \lambda_k, \mu_k, \nu_k)$. One way to obtain a feasible tuple is to set the left side of (16) equal to the zero matrix. This shows (10) is feasible with the following parameter choices.

$$\mu_k = \alpha_k^2(1+\zeta_k^{-1}) \tag{17a}$$

$$\lambda_k = -\alpha_k^2(1+\zeta_k)(1+\delta^2\zeta_k^{-1})M_{22}^{-1} \tag{17b}$$

$$\nu_k = \alpha_k(1+\zeta_k)-\lambda_k M_{21} \tag{17c}$$

$$\rho_k^2 = 1-2\nu_k m+\lambda_k M_{11}+\zeta_k \tag{17d}$$

Since we always have $M_{22} = -1$ in Table 1, it is straightforward to verify that (17) is equivalent to (13) and (14). Notice that we directly have $\mu_k \geq 0$ and $\lambda_k \geq 0$ because $\zeta_k > 0$. In order to ensure $\rho_k^2 \geq 0$ and $\nu_k \geq 0$, we must have $1-2m\alpha_k+\tilde{M}\alpha_k^2(1+\delta^2\zeta_k^{-1}) \geq 0$ and $\alpha_k M_{21}(1+\delta^2\zeta_k^{-1}) \leq 1$, respectively. The first inequality always holds because $\tilde{M} \geq m^2$ and we have $1-2m\alpha_k+\tilde{M}\alpha_k^2(1+\delta^2\zeta_k^{-1}) \geq 1-2m\alpha_k+m^2\alpha_k^2 \geq (1-m\alpha_k)^2 \geq 0$. Based on the conditions $0 \leq \alpha_k M_{21} < 1$ and $\zeta_k \geq \frac{\alpha_k M_{21}\delta^2}{1-\alpha_k M_{21}}$, we conclude that the second inequality always holds as well. Since we have constructed a feasible solution to the LMI (10), the bound (11) follows from Theorem 1. □

Given α_k, Corollary 1 provides a one-dimensional family of solutions to the LMI (10). These solutions are given by (13) and (14) and are parameterized by the auxiliary variable ζ_k. Corollary 1 does not require $\rho_k \leq 1$. Hence it actually does not impose any upper bound on α_k in Case II. Later we will impose refined upper bounds on α_k such that the bound (11) can be transformed into a useful bound in the form of (7). We also want to mention that the stepsize bounds in the above corollary are consistent with the existing results in the machine learning literature. For example, for Case III, the stepsize bound for the standard SGD method is known to be $1/L$ (see Theorem 2.1 in [24]).

Remark 2 We can use (11) to obtain various types of convergence results. For example, when a constant stepsize is used, i.e. $\alpha_k = \alpha$ for all k, a naive analysis can be performed by setting $\zeta_k = \zeta$ for all k. In this case, $(\rho_k, \nu_k, \mu_k, \lambda_k)$ are set to be constants $(\rho, \nu, \mu, \lambda)$. Then, (10) and (11) become independent of k. We can rewrite (11) as

$$\mathbb{E}\|x_{k+1} - x_\star\|^2 \le \rho^2\, \mathbb{E}\|x_k - x_\star\|^2 + (2\lambda G^2 + \mu c^2). \tag{18}$$

If $\rho < 1$, then we may recurse (18) to obtain the following convergence result:

$$\begin{aligned}\mathbb{E}\|x_k - x_\star\|^2 &\le \rho^{2k}\, \mathbb{E}\|x_0 - x_\star\|^2 + \left(\sum_{i=0}^{k-1} \rho^{2i}\right)\left(2\lambda G^2 + \mu c^2\right)\\ &\le \rho^{2k}\, \mathbb{E}\|x_0 - x_\star\|^2 + \frac{2\lambda G^2 + \mu c^2}{1-\rho^2}.\end{aligned} \tag{19}$$

The inequality (19) is an error bound of the familiar form (7). Nevertheless, this bound may be conservative even in the constant stepsize case. To minimize the right-hand side of (11), the objective function for the semidefinite program (10) at step k should be chosen as $\rho_k^2\, \mathbb{E}\|x_k - x_\star\|^2 + (2\lambda_k G^2 + \mu_k c^2)$. Consequently, setting ζ_k to be a constant may introduce conservatism even in the constant stepsize case. To overcome this issue, we will introduce a sequential minimization approach next.

2.2 Sequential minimization approach for biased SGD

We will quantify the convergence behaviors of biased SGD by providing upper bounds for $\mathbb{E}\|x_k - x_\star\|^2$. To do so, we will recursively make use of the bound (11). Suppose δ, c, and G are constant. Define $\mathcal{T}_k \subseteq \mathbb{R}_+^4$ to be the set of tuples $(\rho_k, \lambda_k, \mu_k, \nu_k)$ that are feasible points for the LMI (10). Also define the real number sequence $\{U_k\}_{k\ge 0}$ via the recursion:

$$U_0 \ge \mathbb{E}\,\|x_0 - x_\star\|^2 \quad \text{and} \quad U_{k+1} = \rho_k^2 U_k + 2\lambda_k G^2 + \mu_k c^2 \tag{20}$$

where $(\rho_k, \lambda_k, \mu_k, \nu_k) \in \mathcal{T}_k$. By induction, we can show that U_k provides an upper bound for the error at timestep k. Indeed, if $\mathbb{E}\,\|x_k - x_\star\|^2 \le U_k$, then by Theorem 1, we have $\mathbb{E}\,\|x_{k+1} - x_\star\|^2 \le \rho_k^2\, \mathbb{E}\|x_k - x_\star\|^2 + 2\lambda_k G^2 + \mu_k c^2 \le \rho_k^2 U_k + 2\lambda_k G^2 + \mu_k c^2 = U_{k+1}$. A key issue in computing a useful upper bound U_k is how to choose the tuple $(\rho_k, \lambda_k, \mu_k, \nu_k) \in \mathcal{T}_k$. If the stepsize is constant ($\alpha_k = \alpha$), then $\mathcal{T}_k$ is independent of k. Thus we may choose the same particular solution $(\rho, \lambda, \mu, \nu)$ for each k. Then, based on (19), if $\rho < 1$ we can obtain a bound of the following form for biased SGD:

$$\mathbb{E}\|x_k - x_\star\|^2 \le \rho^{2k} U_0 + \frac{2\lambda G^2 + \mu c^2}{1-\rho^2}. \tag{21}$$

As discussed in Remark 2, the above bound may be unnecessarily conservative. Because of the recursive definition (20), the bound U_k depends solely on U_0 and the parameters $\{\rho_t, \lambda_t, \mu_t\}_{t=0}^{k-1}$. So we can seek the smallest possible upper bound by solving the optimization problem:

$$\begin{aligned} U_{k+1}^{\text{opt}} = \underset{\{\rho_t,\lambda_t,\mu_t,\nu_t\}_{t=0}^{k}}{\text{minimize}} \quad & U_{k+1} \\ \text{subject to} \quad & U_{t+1} = \rho_t^2 U_t + 2\lambda_t G^2 + \mu_t c^2 \qquad 0 \le t \le k \\ & (\rho_t, \lambda_t, \mu_t, \nu_t) \in \mathcal{T}_t \qquad\qquad\qquad\; 0 \le t \le k \end{aligned}$$

A useful fact is that the above optimization problem can be solved in a sequential manner. Formally, we have

Proposition 1 *The following holds for all k.*

$$U_{k+1}^{\text{opt}} = \underset{(\rho,\lambda,\mu,\nu)\in\mathcal{T}_k}{\text{minimize}} \quad \rho^2 U_k^{\text{opt}} + 2\lambda G^2 + \mu c^2 \tag{22}$$

Consequently, a greedy approach where U_{t+1} is optimized in terms of U_t recursively for $t = 0, \ldots, k-1$ yields a bound U_k that is in fact globally optimal over all possible choices of parameters $\{\rho_t, \lambda_t, \mu_t, \nu_t\}_{t=0}^{k}$.

Proof This optimization problem being considered is similar to a dynamic programming and a recursive solution reminiscent of the Bellman equation can be derived for the optimal bound U_k.

$$\begin{aligned}
&U_{k+1}^{\text{opt}} \\
&= \underset{(\rho,\lambda,\mu,\nu)\in\mathcal{T}_k}{\text{minimize}} \left\{ \begin{array}{lll} \underset{\{\rho_t,\lambda_t,\mu_t,\nu_t\}_{t=0}^{k-1}}{\text{minimize}} & U_{k+1} & \\ \text{subject to} & U_{t+1} = \rho_t^2 U_t + 2\lambda_t G^2 + \mu_t c^2 & 0 \le t \le k \\ & (\rho_t, \lambda_t, \mu_t, \nu_t) \in \mathcal{T}_t & 0 \le t < k \\ & (\rho, \lambda, \mu, \nu) = (\rho_k, \lambda_k, \mu_k, \nu_k) & \end{array} \right\} \\
&= \underset{(\rho,\lambda,\mu,\nu)\in\mathcal{T}_k}{\text{minimize}} \left\{ \begin{array}{lll} \underset{\{\rho_t,\lambda_t,\mu_t,\nu_t\}_{t=0}^{k-1}}{\text{minimize}} & \rho^2 U_k + 2\lambda G^2 + \mu c^2 & \\ \text{subject to} & U_{t+1} = \rho_t^2 U_t + 2\lambda_t G^2 + \mu_t c^2 & 0 \le t < k \\ & (\rho_t, \lambda_t, \mu_t, \nu_t) \in \mathcal{T}_t & 0 \le t < k \end{array} \right\} \\
&= \underset{(\rho,\lambda,\mu,\nu)\in\mathcal{T}_k}{\text{minimize}} \quad \rho^2 U_k^{\text{opt}} + 2\lambda G^2 + \mu c^2
\end{aligned} \tag{23}$$

Where the final equality in (23) relies on the fact that $\rho^2 \ge 0$. □

Obtaining an explicit analytical formula for U_k^{opt} is not straightforward, since it involves solving a sequence of semidefinite programs. However, we can make use of Corollary 1 to further upper-bound U_k^{opt}. This works because Corollary 1 gives an analytical parameterization of a *subset* of $\mathcal{T}_k$. Denote this new upper bound by $\hat{U}_k$. By Corollary 1, we have:

$$\begin{aligned}
\hat{U}_{k+1} \;=\; \underset{\zeta>0}{\text{minimize}} \quad & \rho^2 \hat{U}_k + 2\lambda G^2 + \mu c^2 \\
\text{subject to} \quad & \mu = \alpha_k^2 (1 + \zeta^{-1}) \\
& \lambda = \alpha_k^2 (1 + \zeta)(1 + \delta^2 \zeta^{-1}) \\
& \rho^2 = (1 + \zeta)(1 - 2m\alpha_k + \tilde{M}\alpha_k^2 (1 + \delta^2 \zeta^{-1})) \\
& \zeta \ge \tfrac{\alpha_k M_{21} \delta^2}{1 - \alpha_k M_{21}}
\end{aligned} \tag{24}$$

Note that Corollary 1 also places bounds on α_k, which we assume are being satisfied here. The optimization problem (24) is a single-variable smooth constrained problem. It is straightforward to verify that μ, λ, and ρ^2 are convex functions of ζ when $\zeta > 0$. Moreover, the inequality constraint on ζ is linear, so we deduce that (24) is a convex optimization problem.

Thus, we have reduced the problem of recursively solving semidefinite programs (finding U_k^{opt}) to recursively solving single-variable convex optimization problems (finding $\hat{U}_k$). Ultimately, we obtain an upper bound on the expected error of biased SGD that is easy to compute:

$$\mathbb{E}\,\|x_k - x_\star\|^2 \le U_k^{\text{opt}} \le \hat{U}_k \tag{25}$$

Preliminary numerical simulations suggest that $\hat{U}_k$ seems to be equal to U_k^{opt} under the four sets of assumptions in this paper. However, we are unable to show $\hat{U}_k = U_k^{\text{opt}}$ analytically. In the subsequent sections, we will solve the recursion for $\hat{U}_k$ analytically and derive convergence bounds for biased SGD.

2.3 Analytical recursive bounds for biased SGD

We showed in the previous section that $\mathbb{E}\|x_k - x_\star\|^2 \le \hat{U}_k$ for biased SGD, where $\hat{U}_k$ is the solution to (24). We now derive an analytical recursive formula for $\hat{U}_k$. Let us simplify the optimization problem (24). Eliminating ρ, λ, μ, we obtain

$$\begin{aligned}
\hat{U}_{k+1} \;=\; \underset{\zeta>0}{\text{minimize}} \quad & a_k(1+\zeta^{-1}) + b_k(1+\zeta) \\
\text{subject to} \quad & a_k = \alpha_k^2\left(c^2 + 2\delta^2 G^2 + \tilde{M}\delta^2 \hat{U}_k\right) \\
& b_k = \left(1 - 2m\alpha_k + \tilde{M}\alpha_k^2\right)\hat{U}_k + 2\alpha_k^2 G^2 \\
& \zeta \ge \tfrac{\alpha_k M_{21}\delta^2}{1-\alpha_k M_{21}}
\end{aligned} \tag{26}$$

The assumptions on α_k from Corollary 1 also imply that $a_k \ge 0$ and $b_k \ge 0$. We may now solve this problem explicitly and we summarize the solution to (26) in the following lemma.

Lemma 1 *Consider biased SGD (3) with $g \in \mathcal{S}(m,\infty)$ for some $m > 0$, and let $x_\star$ be the unique global minimizer of g. Given one of the four conditions on f_i and the corresponding $M = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22}\end{bmatrix}$ and G from Table 1, further assume α_k is strictly positive and satisfies $M_{21}\alpha_k \le 1$. Then the error bound $\hat{U}_k$ defined in (26) can be computed recursively as follows.*

$$\hat{U}_{k+1} = \begin{cases} \left(\sqrt{a_k} + \sqrt{b_k}\right)^2 & \sqrt{\frac{a_k}{b_k}} \ge \frac{\alpha_k M_{21}\delta^2}{1-\alpha_k M_{21}} \\ a_k + b_k + a_k \frac{1-\alpha_k M_{21}}{\alpha_k M_{21}\delta^2} + b_k \frac{\alpha_k M_{21}\delta^2}{1-\alpha_k M_{21}} & \textit{otherwise} \end{cases} \tag{27}$$

where a_k and b_k are defined in (26)*. We may initialize the recursion at any* $\hat{U}_0 \geq \mathbb{E}\,\|x_0 - x_\star\|^2$.

Proof In Case II, we have $M_{21} = 0$ so the constraint on ζ is vacuously true. Therefore, the only constraint on ζ in (26) is $\zeta > 0$ and we can solve the problem by setting the derivative of the objective function with respect to ζ equal to zero. The result is $\zeta_k = \sqrt{\frac{a_k}{b_k}}$. In Cases I, III, and IV, we have $M_{21} > 0$. By convexity, the optimal ζ_k is either the unconstrained optimum (if it is feasible) or the boundary point (otherwise). Hence (27) holds as desired. Note that if $\delta = c = 0$, then $a_k = 0$. This corresponds to the pathological case where the objective reduces to $b_k(1+\zeta)$. Here, the optimum is achieved as $\zeta \to 0$, which corresponds to $\mu \to \infty$ in (24). This does not cause a problem because $c = 0$ so μ does not appear in the objective function. The recursion (27) then simplifies to $\hat{U}_{k+1} = b_k$. □

Remark 3 If $M_{21} = 0$ (Case II in Table 1) or if $\delta = 0$ (no multiplicative noise), the optimization problem (26) reduces to an unconstrained optimization problem whose solution is

$$\begin{aligned}\hat{U}_{k+1} &= \left(\sqrt{a_k} + \sqrt{b_k}\right)^2 \\ &= \left(\alpha_k\sqrt{c^2 + 2\delta^2 G^2 + \tilde{M}\delta^2\hat{U}_k} + \sqrt{\left(1 - 2m\alpha_k + \tilde{M}\alpha_k^2\right)\hat{U}_k + 2G^2\alpha_k^2}\right)^2\end{aligned} \tag{28}$$

3 Analytical rate bounds for the constant stepsize case

In this section, we present non-recursive error bounds for biased SGD with constant stepsize. Specifically, we assume $\alpha_k = \alpha$ for all k and we either apply Lemma 1 or carefully choose a constant ζ in order to obtain a tractable bound for $\hat{U}_k$. The bounds derived in this section highlight the trade-offs inherent in the design of biased SGD.

3.1 Linearization of the nonlinear recursion

This first result applies to the case where $\delta = 0$ or $M_{21} = 0$ (Case II) and leverages Remark 3 to obtain a bound for biased SGD.

Corollary 2 *Consider biased SGD* (3) *with $g \in \mathcal{S}(m, \infty)$ for some $m > 0$, and let $x_\star$ be the unique global minimizer of g. Given one of the four conditions on f_i and the corresponding $M = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22}\end{bmatrix}$ and G from Table* 1*, further assume that $\alpha_k = \alpha > 0$ (constant stepsize), $M_{21}\alpha \leq 1$, and either $\delta = 0$ or $M_{21} = 0$. Define $p, q, r, s \geq 0$ as follows.*

$$p = \tilde{M}\delta^2\alpha^2, \quad q = (c^2 + 2G^2\delta^2)\alpha^2, \quad r = 1 - 2m\alpha + \tilde{M}\alpha^2, \quad s = 2G^2\alpha^2. \tag{29}$$

Where $\tilde{M} = M_{11} + 2mM_{12}$. *If* $\sqrt{p} + \sqrt{r} < 1$ *then we have the following iterate error bound:*

$$\mathbb{E}\,\|x_k - x_\star\|^2 \le \left(\frac{p\sqrt{\hat{U}_\star}}{\sqrt{p\hat{U}_\star + q}} + \frac{r\sqrt{\hat{U}_\star}}{\sqrt{r\hat{U}_\star + s}} \right)^k \mathbb{E}\,\|x_0 - x_\star\|^2 + \hat{U}_\star, \tag{30}$$

where the fixed point $\hat{U}_\star$ *is given by*

$$\hat{U}_\star = \frac{(p-r)(s-q) + q + s + 2\sqrt{ps^2 + q^2 r + qs(1-p-r)}}{(p-r)^2 - 2(p+r) + 1}. \tag{31}$$

Proof By Remark 3, we have the nonlinear recursion (28) for $\hat{U}_k$. This recursion is of the form

$$\hat{U}_{k+1} = \left(\sqrt{p\hat{U}_k + q} + \sqrt{r\hat{U}_k + s} \right)^2, \tag{32}$$

where $p, q, r, s > 0$ are given in (29). It is straightforward to verify that the right-hand side of (32) is a monotonically increasing concave function of $\hat{U}_k$ and its asymptote is a line of slope $(\sqrt{p}+\sqrt{r})^2$. Thus, (32) will have a unique fixed point when $\sqrt{p}+\sqrt{r} < 1$. We will return to this condition shortly. When a fixed point exists, it is found by setting $\hat{U}_k = \hat{U}_{k+1} = \hat{U}_\star$ in (32) and yields $U_\star$ given by (31). The concavity property further guarantees that any first-order Taylor expansion of the right-hand side of (32) yields an upper bound to $\hat{U}_{k+1}$. Expanding about $\hat{U}_\star$, we obtain:

$$\hat{U}_{k+1} - \hat{U}_\star \le \left(\frac{p\sqrt{\hat{U}_\star}}{\sqrt{p\hat{U}_\star + q}} + \frac{r\sqrt{\hat{U}_\star}}{\sqrt{r\hat{U}_\star + s}} \right) \left(\hat{U}_k - \hat{U}_\star \right) \tag{33}$$

which leads to the following non-recursive bound for biased SGD.

$$\begin{aligned} \mathbb{E}\,\|x_k - x_\star\|^2 \le \hat{U}_k &\le \left(\frac{p\sqrt{\hat{U}_\star}}{\sqrt{p\hat{U}_\star + q}} + \frac{r\sqrt{\hat{U}_\star}}{\sqrt{r\hat{U}_\star + s}} \right)^k (\hat{U}_0 - \hat{U}_\star) + \hat{U}_\star \\ &\le \left(\frac{p\sqrt{\hat{U}_\star}}{\sqrt{p\hat{U}_\star + q}} + \frac{r\sqrt{\hat{U}_\star}}{\sqrt{r\hat{U}_\star + s}} \right)^k \hat{U}_0 + \hat{U}_\star \end{aligned} \tag{34}$$

Since this bound holds for any $\hat{U}_0 \ge \mathbb{E}\,\|x_0 - x_\star\|^2$, it holds in particular when we have equality, and thus we obtain (30) as required. □

The condition that $\sqrt{p} + \sqrt{r} < 1$ from Corollary 2, which is necessary for the existence of a fixed-point of (32), is equivalent to an upper bound on α. After manipulation, it amounts to:

$$\alpha < \frac{2(m - \delta\sqrt{\tilde{M}})}{\tilde{M}(1-\delta^2)} \tag{35}$$

Therefore, we can ensure that $\sqrt{p}+\sqrt{r}<1$ when $\delta < m/\sqrt{\tilde{M}}$, and α is sufficiently small. If $\delta=0$, the stepsize bound (35) is only relevant in Case II. For Cases I, III, and IV, the bound $M_{21}\alpha \le 1$ imposes a stronger restriction on α (see Corollary 1). If $\delta \neq 0$, we only consider Case II ($M_{21}=0$) and the resultant bound for α is $\frac{m-\sqrt{2}L\delta}{L^2(1-\delta^2)}$. The condition $\delta < m/\sqrt{\tilde{M}}$ becomes $\delta < m/(\sqrt{2}L)$.

To see the trade-offs in the design of biased SGD, we can take Taylor expansions of several key quantities about $\alpha=0$ to see how changes in α affect convergence:

$$\hat{U}_\star \approx \frac{c^2+2\delta^2G^2}{m^2-\delta^2\tilde{M}} + \frac{m\left(c^2(\tilde{M}-m^2)+2\left(1-\delta^2\right)G^2m^2\right)}{(m^2-\delta^2\tilde{M})^2}\alpha + O(\alpha^2) \tag{36a}$$

$$\left(\frac{p\sqrt{\hat{U}_\star}}{\sqrt{p\hat{U}_\star+q}}+\frac{r\sqrt{\hat{U}_\star}}{\sqrt{r\hat{U}_\star+s}}\right) \approx 1-\frac{(m^2-\delta^2\tilde{M})}{m}\alpha + O(\alpha^2) \tag{36b}$$

We conclude that when $\delta < m/\sqrt{\tilde{M}}$, biased SGD converges linearly to a ball whose radius is roughly $\hat{U}_\star \ge 0$. One can decrease the stepsize α to control the final error $\hat{U}_\star$. However, due to the errors in the individual gradient updates, one cannot guarantee the final error $\mathbb{E}\,\|x_k-x_\star\|^2$ smaller than $\frac{c^2+2\delta^2G^2}{m^2-\delta^2\tilde{M}}$. This is consistent with our intuition; one could inject noise in an adversarial manner to shift the optimum point away from $x_\star$ so there is no way to guarantee that $\{x_k\}$ converges to $x_\star$ just by decreasing the stepsize α.

Remark 4 One can check that the left side of (36b) is not differentiable at $(c,\alpha)=(0,0)$. Consequently, taking a Taylor expansion with respect to α and then setting $c=0$ does not yield the same result as first setting $c=0$ and then taking a Taylor expansion with respect to α of the resulting expression. This explains why (36b) does not reduce to $\rho^2=1-2m\alpha+O(\alpha^2)$ when $c=\delta=0$. It is worth noting that the higher order term $O(\alpha^2)$ in (36b) depends on c. Indeed, it blows up as $c\to 0$. Therefore, the rate formula (36b) only describes the stepsize design trade-off for a fixed positive c and sufficiently small α. Similar situation even holds for the case where $G=0$. As long as $c\neq 0$, the rate formula is not going to reduce to $\rho^2=1-2m\alpha+O(\alpha^2)$ due to the fact that the left side of (36b) is not differentiable at $(c,\alpha)=(0,0)$.

The non-recursive bound (30) relied on a linearization of the recursive formula (32), which involved a time-varying ζ_k. It is emphasized that we assumed that either $\delta=0$ or $M_{21}=0$. In the other cases, namely $\delta>0$ and $M_{21}>0$ (Case I, III, or IV), we cannot ignore the additional condition $\zeta_k \ge \frac{\alpha M_{21}\delta^2}{1-\alpha M_{21}}$ and we must use the hybrid recursive formula (27). This hybrid formulation is more problematic to solve explicitly. However, if we are mostly interested in the regime where α is small, we can obtain non-recursive bounds similar to (30) by carefully choosing a constant ζ for all k. We will develop these bounds in the next section.

3.2 Non-recursive bounds via a fixed ζ parameter

When α is small, we can choose $\zeta=m\alpha$ and we obtain the following result.

Corollary 3 *Consider biased SGD (3) with $g \in \mathcal{S}(m,\infty)$ for some $m > 0$, and let $x_\star$ be the unique global minimizer of g. Given one of the four conditions on f_i and the corresponding $M = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix}$ and G from Table 1, further assume that $\alpha_k = \alpha > 0$ (constant stepsize), and $M_{21}\left(\alpha + \frac{\delta^2}{m}\right) \le 1$.[5] Finally, assume that*

$$0 < \tilde{\rho}^2 < 1 \quad \textit{where } \tilde{\rho}^2 = 1 - \tfrac{m^2 - \tilde{M}\delta^2}{m}\alpha + (\tilde{M}(1+\delta^2) - 2m^2)\alpha^2 + \tilde{M}m\alpha^3. \tag{37}$$

Note (37) holds for α sufficiently small. Then, we have the following error bound for the iterates

$$\mathbb{E}\|x_k - x_\star\|^2 \le \tilde{\rho}^{2k}\,\mathbb{E}\,\|x_0 - x_\star\|^2 + \tilde{U}_\star \tag{38}$$

where $\tilde{U}_\star$ is given by

$$\tilde{U}_\star = \frac{2\delta^2 G^2 + c^2 + m(c^2 + 2G^2(1+\delta^2))\alpha + 2m^2G^2\alpha^2}{(m^2 - \tilde{M}\delta^2) - m(\tilde{M}(1+\delta^2) - 2m^2)\alpha - \tilde{M}m^2\alpha^2} \tag{39}$$

Proof Set $\zeta = m\alpha$ in the optimization problem (26). This defines a new recursion for a quantity $\tilde{U}_k$ that upper-bounds $\hat{U}_k$ since we are choosing a possibly sub-optimal ζ. Our assumption $M_{21}\left(\alpha + \frac{\delta^2}{m}\right) \le 1$ guarantees that $\zeta \ge \frac{\alpha M_{21}\delta^2}{1-\alpha M_{21}}$ when $\zeta = m\alpha$. Hence our choice of ζ is a feasible choice for (26). This leads to:

$$\begin{aligned} \tilde{U}_{k+1} &= a_k(1 + \tfrac{1}{m\alpha}) + b_k(1 + m\alpha) \\ &= \tilde{\rho}^2\tilde{U}_k + \left(\alpha^2(c^2 + 2\delta^2G^2)(1 + \tfrac{1}{m\alpha}) + 2\alpha^2G^2(1 + m\alpha)\right) \end{aligned}$$

This is a simple linear recursion that we can solve explicitly in a similar way to the recursion in Remark 2. After simplifications, we obtain (38) and (39). □

The linear rate of convergence in (38) is of the same order as the one obtained in Corollary 2 and (36b). Namely,

$$\tilde{\rho}^2 \approx 1 - \frac{(m^2 - \tilde{M}\delta^2)}{m}\alpha + O(\alpha^2) \tag{40}$$

Likewise, the limiting error $\tilde{U}_\star$ from (39) can be expanded as a series in α and we obtain a result that matches the small-α limit of $\hat{U}_\star$ from (36a) up to linear terms. Namely,

$$\tilde{U}_\star \approx \frac{c^2 + 2\delta^2G^2}{m^2 - \tilde{M}\delta^2} + \frac{m\left(c^2(\tilde{M} - m^2) + 2\left(1 - \delta^2\right)G^2m^2\right)}{(m^2 - \delta^2\tilde{M})^2}\alpha + O(\alpha^2) \tag{41}$$

[5] When $M_{21} = 0$, this condition always holds. When $\delta = 0$, this condition is equivalent to $M_{21}\alpha \le 1$. Hence the above corollary can be directly applied if $M_{21} = 0$ or $\delta = 0$. If $M_{21} > 0$ and $\delta > 0$, the condition $M_{21}\left(\alpha + \frac{\delta^2}{m}\right) \le 1$ can be rewritten as a condition on α in a case-by-case manner.

Table 2 Upper bound on δ for the four different cases described in Table 1

Case	I	II	III	IV
$\tilde{M} = M_{11} + 2mM_{12}$	m^2	$2L^2$	$2mL$	$2m^2$
δ bound	1	$\frac{m}{\sqrt{2}L}$	$\sqrt{\frac{m}{2L}}$	$\sqrt{\frac{m}{L+m}}$

Therefore, (38) can give a reasonable non-recursive bound for biased SGD with small α even for the cases where $M_{21} > 0$ and $\delta > 0$.

Now we discuss the acceptable relative noise level under various assumptions on f_i. Based on (40), we need $m^2 - \tilde{M}\delta^2 > 0$ to ensure $\tilde{\rho}^2 < 1$ for sufficiently small α. The other constraint $M_{21}\left(\alpha + \frac{\delta^2}{m}\right) \leq 1$ enforces $M_{21}\delta^2 < m$. Depending on which case we are dealing with, the conditions $\delta < m/\sqrt{\tilde{M}}$ and $M_{21}\delta^2 < m$ impose an upper bound on admissible values of δ. See Table 2.

We can clearly see that for ℓ_2-regularized logistic regression and support vector machines which admit the assumption in Case I, biased SGD is robust to the relative noise. Given the condition $\delta < 1$, the iterates of biased SGD will stay in some ball, although the size of the ball could be large. Comparing the bound for Cases II, III, and IV, we can see the allowable relative noise level increases as the assumptions on f_i become stronger.

As previously mentioned, the bound of Corollary 3 requires a sufficiently small α. Specifically, the stepsize α must satisfy $M_{21}\left(\alpha + \frac{\delta^2}{m}\right) \leq 1$ and (37), which can be solved to obtain explicit upper bounds on α. Details are omitted.

Sensitivity analysis. Based on (40), the convergence rate $\bar{\rho}^2$ can be estimated as $1 - \frac{m^2 - \tilde{M}\delta^2}{m}\alpha$ for small α, which is independent of c. Hence the misspecification in the value of c does not impact the value of $\bar{\rho}^2$. The derivative of $1 - \frac{m^2 - \tilde{M}\delta^2}{m}\alpha$ with respect to δ is $2\tilde{M}\delta\alpha/m$. Hence, if we perturb the value of δ by ϵ, the change in the value of $\bar{\rho}^2$ is roughly equal to $2\tilde{M}\delta\alpha\epsilon/m$. Similarly, we can perform a sensitivity analysis for the final optimization error term $\bar{U}_\star$ by taking the derivative of the right side of (41) with respect to δ (or c).

Conservatism of Corollary 3. Corollary 3 gives a reasonable non-recursive bound for biased SGD with small α. However, if we consider Case II, it can be much more conservative than Corollary 2 for relatively larger α . We use a numerical example to illustrate this. Consider $m = 1$, $L = 100$, $G = 5$, and $c = 1$. We set $\delta = 0.0021 < 0.0071 = m/\sqrt{\tilde{M}}$. Based on (35), we know Corollary 2 works for $\alpha < 7 \times 10^{-5}$. Based on (37), we can show Corollary 3 works for $\alpha < 4.55 \times 10^{-5}$. Obviously, Corollary 2 works for a larger range of α. By numerical simulations, it is straightforward to verify that $\bar{U}_\star \to \infty$ and $\bar{\rho} \to 1$ if we apply Corollary 3 to the case where $\alpha = 4.56 \times 10^{-5}$. In contrast, if we apply Corollary 2 to the case where $\alpha = 4.56 \times 10^{-5}$, we can obtain $\hat{U}_\star = 4.8207$. The associated convergence rate is $1 - 1.74 \times 10^{-5}$. Clearly, Corollary 2 gives a much more reasonable bound in this case. We have tried different problem parameters and have observed similar trends. In general, for Case II, Corollary 3 is more conservative than Corollary 2 if relatively large α is considered. The advantage of Corollary 3 is that it is general enough to cover Cases I, III, and IV.

4 Further discussion

4.1 Connections to existing SGD results

In this section, we relate the results of Theorem 1 and its corollaries to existing results on standard SGD. We also discuss the effect of replacing our error model (4) with IID noise.

If there is no noise at all, $c = \delta = 0$ and none of the approximations of Sect. 3 are required to obtain an analytical bound on the iteration error. Returning to Theorem 1 and Corollary 1, the objective to be minimized no longer depends on μ_k. Examining (13), we conclude that optimality occurs as $\zeta \to 0$ ($\mu \to \infty$). This leads directly to the bound

$$\mathbb{E}\,\|x_{k+1} - x_\star\|^2 \leq (1 - 2m\alpha_k + \tilde{M}\alpha_k^2)\,\mathbb{E}\,\|x_k - x_\star\|^2 + 2G^2\alpha_k^2, \tag{42}$$

where α_k is constrained such that $M_{21}\alpha_k \leq 1$. The bound (42) directly leads to existing convergence results for standard SGD. For example, we can apply the argument in Remark 2 to obtain the following bound for standard SGD with a constant stepsize $\alpha_k = \alpha$

$$\mathbb{E}\|x_k - x_\star\|^2 \leq \left(1 - 2m\alpha + \tilde{M}\alpha^2\right)^k \mathbb{E}\|x_0 - x_\star\|^2 + \frac{2G^2\alpha}{2m - \tilde{M}\alpha}, \tag{43}$$

where α is further required to satisfy $1 - 2m\alpha + \tilde{M}\alpha^2 \leq 1$. For Cases I, III, and IV, the condition $M_{21}\alpha \leq 1$ dominates, and the valid values of α are documented in Corollary 1. For Case II, the condition $\alpha \leq 2m/\tilde{M}$ dominates and the upper bound on α is m/L^2.

The bound recovers existing results that describe the design trade-off of standard (noiseless) SGD under a variety of conditions [22–24]. Case I is a slight variant of the well-known result [23, Prop. 3.4]. The extra factor of 2 in the rate and errors terms are due to the fact that [23, Prop. 3.4] poses slightly different conditions on g and f_i. Cases II and III are also well-known [15,22,24].

Remark 5 If the error term e_k is IID noise with zero mean and bounded variance, then a slight modification to our analysis yields the bound

$$\mathbb{E}\,\|x_{k+1} - x_\star\|^2 \leq (1 - 2m\alpha_k + \tilde{M}\alpha_k^2)\,\mathbb{E}\,\|x_k - x_\star\|^2 + (2G^2 + \sigma^2)\alpha_k^2, \tag{44}$$

where $\sigma^2 \geq \mathbb{E}\,\|e_k\|^2$. The detailed proof is omitted.

4.2 Adaptive stepsize via sequential minimization

In Sect. 3, we fixed $\alpha_k = \alpha$ and derived bounds on the worst-case performance of biased SGD. In this section, we discuss the potential impacts of adopting time-varying stepsizes. First, we refine the bounds by optimizing over α_k as well. What makes this approach tractable is that in Theorem 1, the LMI (10) is also linear in α_k. Therefore, we can easily include α_k as one of our optimization variables.

In fact, the development of Sect. 2.2 carries through if we augment the set $\mathcal{T}_k$ to be the set of tuples $(\rho_k, \lambda_k, \mu_k, \nu_k, \alpha_k)$ that makes the LMI (10) feasible. We then obtain a Bellman-like equation analogous to (23) that holds when we also optimize over α at every step. The net result is an optimization problem similar to (26) but that now includes α as a variable:

$$\begin{aligned} V_{k+1} = \underset{\alpha>0,\,\zeta>0}{\text{minimize}} \quad & a_k(1+\zeta^{-1}) + b_k(1+\zeta) \\ \text{subject to} \quad & a_k = \alpha^2\left(c^2 + 2\delta^2 G^2 + \tilde{M}\delta^2 V_k\right) \\ & b_k = \left(1 - 2m\alpha + \tilde{M}\alpha^2\right) V_k + 2\alpha^2 G^2 \\ & \alpha M_{21}(1+\delta^2\zeta^{-1}) \le 1 \end{aligned} \tag{45}$$

As we did in Sect. 2.2, we can show that $\mathbb{E}\,\|x_k - x_\star\|^2 \le V_k$ for any iterates of biased SGD. We would like to learn two things from (45): how the optimal α changes as a function of k in order to produce the fastest possible convergence rate, and whether this optimized rate is different from the rate we obtained when assuming α was constant in Sect. 3.

To simplify the analysis, we will restrict our attention to Case II, where $M_{21} = 0$ and $\tilde{M} = 2L^2$. In this case, the inequality constraint in (45) is satisfied for any $\alpha > 0$ and $\zeta > 0$, so it may be removed. Observe that the objective in (45) is a quadratic function of α.

$$\begin{aligned} a_k(1+\zeta^{-1}) + b_k(1+\zeta) = (1+\zeta)V_k - 2m(1+\zeta)V_k\alpha + (1+\zeta^{-1})(c^2 + 2G^2\delta^2 \\ + \tilde{M}V_k\delta^2 + 2G^2\zeta + \tilde{M}V_k\zeta)\alpha^2 \end{aligned} \tag{46}$$

This quadratic is always positive definite, and the optimal α is given by:

$$\alpha_k^{\text{opt}} = \frac{mV_k\zeta}{(c^2 + 2\delta^2 G^2 + \delta^2\tilde{M}V_k) + (2G^2 + \tilde{M}V_k)\zeta} \tag{47}$$

Substituting (47) into (45) to eliminate α, we obtain the optimization problem:

$$V_{k+1} \;=\; \underset{\zeta>0}{\text{minimize}} \quad \frac{(\zeta+1)V_k\left(c^2 + (2G^2 + \tilde{M}V_k)(\delta^2+\zeta) - m^2V_k\zeta\right)}{c^2 + \left(2G^2 + \tilde{M}V_k\right)\left(\delta^2 + \zeta\right)} \tag{48}$$

By taking the second derivative with respect to ζ of the objective function in (48), one can check that we will have convexity as long as $(2G^2 + \tilde{M}V_k)(1-\delta^2) \ge c^2$. In other words, we have convexity as long as the noise parameters c and δ are not too large. If this bound holds for $V_k = 0$, then it will hold for any $V_k > 0$. So it suffices to ensure that $2G^2(1-\delta^2) \ge c^2$.

Upon careful analysis of the objective function, we note that when $\zeta = 0$, we obtain $V_{k+1} = V_k$. In order to obtain a decrease for some $\zeta > 0$, we require a negative derivative at $\zeta = 0$. This amounts to the condition: $c^2 + (2G^2 + \tilde{M}V_k)\delta^2 < m^2V_k$. As

V_k gets smaller, this condition will eventually be violated. Specifically, the condition holds whenever $m^2 - \tilde{M}\delta^2 > 0$ and

$$V_k > \frac{c^2 + 2\delta^2 G^2}{m^2 - \tilde{M}\delta^2}$$

Note that this is the same limit as was observed in the constant-α limits $\hat{U}_\star$ and $\tilde{U}_\star$ when $\alpha \to 0$ in (36a) and (41), respectively. This is to be expected; the biased gradient information introduces an uncontrollable bias (which is quantified as $\frac{c^2+2\delta^2 G^2}{m^2-\tilde{M}\delta^2}$) into the final optimization result, and this can not be overcome by any stepsize rules. Notice that we have not ruled out the possibility that V_k suddenly jumps below $\frac{c^2+2\delta^2 G^2}{m^2-\tilde{M}\delta^2}$ at some k and then stays unchanged after that. We will make a formal argument to rule out this possibility in the next lemma. Moreover, the question remains as to whether this minimal error can be achieved *faster* by varying α_k in an optimal manner. We describe the final nonlinear recursion in the next lemma.

Lemma 2 *Consider biased SGD (3) with $g \in \mathcal{S}(m, \infty)$ for some $m > 0$, and let $x_\star$ be the unique global minimizer of g. Suppose Case II holds and (M, G) are the associated values from Table 1. Further assume $2G^2(1-\delta^2) \geq c^2$ and $V_0 > \frac{c^2+2\delta^2 G^2}{m^2-\tilde{M}\delta^2} = V_\star$.*

1. *The sequential optimization problem (48) can be solved using the following nonlinear recursion*

$$V_{k+1} = \frac{V_k}{(2G^2 + \tilde{M}V_k)^2} \times \left(\sqrt{(2G^2 + (\tilde{M} - m^2)V_k)\big((2G^2 + \tilde{M}V_k)(1-\delta^2) - c^2\big)} + \sqrt{m^2 V_k (c^2 + \delta^2(2G^2 + \tilde{M}V_k))}\right)^2 \tag{49}$$

   and V_k satisfies $V_k > V_\star$ for all k.
2. *Suppose $\hat{U}_0 = V_0 \geq \mathbb{E}\,\|x_0 - x_\star\|^2$ (all recurrences are initialized the same way), then $\{V_k\}_{k\geq 0}$ provides an upper bound to the iterate error satisfying $\mathbb{E}\,\|x_k - x_\star\|^2 \leq V_k \leq \hat{U}_k$.*
3. *The sequence $\{V_k\}_{k\geq 0}$ converges to $V_\star$:*

$$\lim_{k\to\infty} V_k = V_\star = \frac{c^2 + 2\delta^2 G^2}{m^2 - \tilde{M}\delta^2}$$

Proof See “Appendix B”. □

To learn more about the rate of convergence, we can once again use a Taylor series approximation. Specializing to Case II (where $\tilde{M} > 0$), we can consider two cases. When V_k is large, perform a Taylor expansion of (49) about $V_k = \infty$ and obtain:

$$V_{k+1} \approx \left(\frac{m\delta + \sqrt{(\tilde{M}-m^2)(1-\delta^2)}}{\tilde{M}}\right) V_k + O(1)$$

In other words, we obtain linear convergence. When V_k is close to $V_\star$, the behavior changes. To see this, perform a Taylor expansion of (49) about $V_k = V_\star$ and obtain:

$$V_{k+1} \approx V_k - \frac{(m^2 - \tilde{M}\delta^2)^3}{4m^2\big(c^2(\tilde{M} - m^2) + 2G^2m^2(1-\delta^2)\big)}(V_k - V_\star)^2 + O((V_k - V_\star)^3) \tag{50}$$

We will ignore the higher-order terms, and apply the next lemma to show that the above recursion roughly converges at a $O(1/k)$ rate.

Lemma 3 *Consider the recurrence relation*

$$v_{k+1} = v_k - \eta v_k^2 \qquad \textit{for } k = 0, 1, \ldots \tag{51}$$

where $v_0 > 0$ *and* $0 < \eta < v_0^{-1}$. *Then the iterates satisfy the following bound for all* $k \geq 0$.

$$v_k \leq \frac{1}{\eta k + v_0^{-1}} \tag{52}$$

Proof The recurrence (51) is equivalent to $\eta v_{k+1} = \eta v_k - (\eta v_k)^2$ with $0 < \eta v_0 < 1$. Clearly, the sequence $\{\eta v_k\}_{k\geq 0}$ is monotonically decreasing to zero. To bound the iterates, invert the recurrence:

$$\frac{1}{\eta v_{k+1}} = \frac{1}{\eta v_k - (\eta v_k)^2} = \frac{1}{\eta v_k} + \frac{1}{1 - \eta v_k} \geq \frac{1}{\eta v_k} + 1$$

Recursing the above inequality, we obtain: $\frac{1}{\eta v_k} \geq \frac{1}{\eta v_0} + k$. Inverting this inequality yields (52), as required. □

Applying Lemma 3 to the sequence $v_k = V_k - V_\star$ defined in (50), we deduce that when V_k is close to its optimal value of $V_\star$, we have:

$$V_k \sim V_\star + \frac{1}{\eta k + (V_0 - V_\star)^{-1}} \quad \text{with: } \eta = \frac{(m^2 - \tilde{M}\delta^2)^3}{4m^2\big(c^2(\tilde{M} - m^2) + 2G^2m^2(1-\delta^2)\big)} \tag{53}$$

We can also examine how α_k changes in this optimal recursive iteration by taking (47) and substituting the optimal ζ found in the optimization of Lemma 2. The result is messy, but a Taylor expansion about $V_k = V_\star$ reveals that

$$\alpha_k^{\text{opt}} \approx \frac{(m^2 - \tilde{M}\delta^2)^2}{2m\big(c^2(\tilde{M} - m^2) + 2G^2m^2(1-\delta^2)\big)}(V_k - V_\star) + O((V_k - V_\star)^2).$$

So when V_k is close to $V_\star$, we should be decreasing α_k to zero at a rate of $O(1/k)$ so that it mirrors the rate at which $V_k - V_\star$ goes to zero in (53).

Finally, we want to mention that calculating α_k^{opt} requires one to know the problem parameters $(m, \tilde{M}, \delta, c, G, V_0)$ in advance. This restricts the applicability of such adaptive stepsize rules for practical problems. Nevertheless, our results in this section bring new theoretical insights for the potential impacts of time-varying stepsizes on the performance of biased SGD. In summary, adopting an optimized time-varying stepsize still roughly yields a rate of $O(1/k)$, which is consistent with the sublinear convergence rate of standard SGD with diminishing stepsize. It is possible that the well-known lower complexity bounds for standard SGD in [1] can be extended to the inexact case, although a formal treatment is beyond the scope of this paper.

Appendix

A Proof of Theorem 1

First notice that since i_k is uniformly distributed on $\{1, \ldots, n\}$ and x_k and i_k are independent, we have:

$$\mathbb{E}\big(u_k \,\big|\, x_k\big) = \mathbb{E}\big(\nabla f_{i_k}(x_k) \,\big|\, x_k\big) = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(x_k) = \nabla g(x_k)$$

Consequently, we have:

$$\mathbb{E}\left(\begin{bmatrix} x_k - x_\star \\ u_k \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} -2mI_p & I_p \\ I_p & 0_p \end{bmatrix} \begin{bmatrix} x_k - x_\star \\ u_k \end{bmatrix} \,\middle|\, x_k \right) = \begin{bmatrix} x_k - x_\star \\ \nabla g(x_k) \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} -2mI_p & I_p \\ I_p & 0_p \end{bmatrix} \begin{bmatrix} x_k - x_\star \\ \nabla g(x_k) \end{bmatrix} \geq 0 \tag{54}$$

where the inequality in (54) follows from the definition of $g \in \mathcal{S}(m, \infty)$.

Next we prove (12), let's start with Case I, the boundedness constraint $\|\nabla f_i(x_k)\| \leq \beta$ implies that $\|u_k\| \leq \beta$ for all k. Rewrite as a quadratic form to obtain:

$$\begin{bmatrix} x_k - x_\star \\ u_k \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} 0_p & 0_p \\ 0_p & -I_p \end{bmatrix} \begin{bmatrix} x_k - x_\star \\ u_k \end{bmatrix} \geq -\beta^2 \tag{55}$$

The boundedness constraint $\|\nabla f_i(x_k) - mx_k\| \leq \beta$ implies that:

$$\begin{aligned} \|u_k - m(x_k - x_\star)\|^2 &\leq \|(u_k - mx_k) + mx_\star\|^2 + \|(u_k - mx_k) - mx_\star\|^2 \\ &= 2\,\|u_k - mx_k\|^2 + 2m^2\,\|x_\star\|^2 \\ &\leq 2\beta^2 + 2m^2\,\|x_\star\|^2 \end{aligned}$$

As in the proof of Case I, rewrite the above inequality as a quadratic form and we obtain the second row of Table 1.

To prove the three remaining cases, we begin by showing that an inequality of the following form holds for each f_i:

$$\begin{bmatrix} x_k - x_\star \\ \nabla f_i(x_k) - \nabla f_i(x_\star) \end{bmatrix}^\mathsf{T} \begin{bmatrix} M_{11} I_p & M_{12} I_p \\ M_{21} I_p & -2I_p \end{bmatrix} \begin{bmatrix} x_k - x_\star \\ \nabla f_i(x_k) - \nabla f_i(x_\star) \end{bmatrix} \geq 0 \tag{56}$$

The verification for (56) follows directly from the definitions of L-smoothness and convexity. In the smooth case (Definition 1), for example, $\|\nabla f_i(x_k) - \nabla f_i(x_\star)\| \leq L\|x_k - x_\star\|$. So (56) holds with $M_{11} = 2L^2$, $M_{12} = M_{21} = 0$. The cases for $\mathcal{F}(0, L)$ and $\mathcal{F}(m, L)$ follow directly from Definition 2. In Table 1, we always have $M_{22} = -1$. Therefore,

$$\begin{aligned} &\mathbb{E}\left(\begin{bmatrix} x_k - x_\star \\ u_k \end{bmatrix}^\mathsf{T} \begin{bmatrix} M_{11} I_p & M_{12} I_p \\ M_{21} I_p & M_{22} I_p \end{bmatrix} \begin{bmatrix} x_k - x_\star \\ u_k \end{bmatrix} \;\middle|\; x_k \right) \\ &\quad = \frac{1}{n}\sum_{i=1}^n \begin{bmatrix} x_k - x_\star \\ \nabla f_i(x_k) \end{bmatrix}^\mathsf{T} \begin{bmatrix} M_{11} I_p & M_{12} I_p \\ M_{21} I_p & 0_p \end{bmatrix} \begin{bmatrix} x_k - x_\star \\ \nabla f_i(x_k) \end{bmatrix} - \frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x_k)\|^2 \end{aligned} \tag{57}$$

Since $\frac{1}{n}\sum_{i=1}^n \nabla f_i(x_\star) = \nabla g(x_\star) = 0$, the first term on the right side of (57) is equal to

$$\frac{1}{n}\sum_{i=1}^n \begin{bmatrix} x_k - x_\star \\ \nabla f_i(x_k) - \nabla f_i(x_\star) \end{bmatrix} \begin{bmatrix} M_{11} I_p & M_{12} I_p \\ M_{21} I_p & 0_p \end{bmatrix} \begin{bmatrix} x_k - x_\star \\ \nabla f_i(x_k) - \nabla f_i(x_\star) \end{bmatrix}$$

Based on the constraint condition (56), we know that the above term is greater than or equal to $\frac{2}{n}\sum_{i=1}^n \|\nabla f_i(x_k) - \nabla f_i(x_\star)\|^2$. Substituting this fact back into (57) leads to the inequality:

$$\begin{aligned} &\mathbb{E}\left(\begin{bmatrix} x_k - x_\star \\ u_k \end{bmatrix}^\mathsf{T} \begin{bmatrix} M_{11} I_p & M_{12} I_p \\ M_{21} I_p & M_{22} I_p \end{bmatrix} \begin{bmatrix} x_k - x_\star \\ u_k \end{bmatrix} \;\middle|\; x_k \right) \\ &\quad \geq \frac{1}{n}\sum_{i=1}^n \left(2\|\nabla f_i(x_k) - \nabla f_i(x_\star)\|^2 - \|\nabla f_i(x_k)\|^2 \right) \\ &\quad = \frac{1}{n}\sum_{i=1}^n \left(\|\nabla f_i(x_k) - 2\nabla f_i(x_\star)\|^2 - 2\|\nabla f_i(x_\star)\|^2 \right) \\ &\quad \geq -\frac{2}{n}\sum_{i=1}^n \|\nabla f_i(x_\star)\|^2 \end{aligned} \tag{58}$$

Taking the expectation of both sides, we arrive at (12), as desired. Now we are ready to prove our main theorem. By Schur complement, (10) is equivalent to (15), which can be further rewritten as

$$\left(\begin{bmatrix} 1-\rho_k^2 & -\alpha_k & -\alpha_k \\ -\alpha_k & \alpha_k^2 & \alpha_k^2 \\ -\alpha_k & \alpha_k^2 & \alpha_k^2 \end{bmatrix} + \nu_k \begin{bmatrix} -2m & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \lambda_k \begin{bmatrix} M_{11} & M_{12} & 0 \\ M_{21} & M_{22} & 0 \\ 0 & 0 & 0 \end{bmatrix} + \mu_k \begin{bmatrix} 0 & 0 & 0 \\ 0 & \delta^2 & 0 \\ 0 & 0 & -1 \end{bmatrix}\right) \otimes I_p \preceq 0 \tag{59}$$

Since $x_{k+1} - x_\star = x_k - x_\star - \alpha_k(u_k + e_k)$, we have

$$\begin{bmatrix} x_k - x_\star \\ u_k \\ e_k \end{bmatrix}^\mathsf{T} \left(\begin{bmatrix} 1 & -\alpha_k & -\alpha_k \\ -\alpha_k & \alpha_k^2 & \alpha_k^2 \\ -\alpha_k & \alpha_k^2 & \alpha_k^2 \end{bmatrix} \otimes I_p\right) \begin{bmatrix} x_k - x_\star \\ u_k \\ e_k \end{bmatrix} = \|x_{k+1} - x_\star\|^2 \tag{60}$$

Now we can left and right multiply (59) by $[(x_k - x_\star)^\mathsf{T}, u_k^\mathsf{T}, e_k^\mathsf{T}]$ and $[(x_k - x_\star)^\mathsf{T}, u_k^\mathsf{T}, e_k^\mathsf{T}]^\mathsf{T}$, and apply the inequalities (4), (54), and (12) to get the desired conclusion. □

B Proof of Lemma 2

We use an induction argument to prove Item 1. For simplicity, we denote (48) as $V_{k+1} = h(V_k)$. Suppose we have $V_k = h(V_{k-1})$ and $V_{k-1} > V_\star$. We are going to show $V_{k+1} = h(V_k)$ and $V_k > V_\star$. We can rewrite (48) as

$$V_{k+1} = \underset{\zeta > 0}{\text{minimize}} \quad A_k(1 + Z_k^{-1}) + B_k(1 + Z_k) \tag{61}$$

where A_k, B_k, and Z_k are defined as

$$\begin{aligned} A_k &= \frac{m^2 V_k^2 \left(c^2 + (2G^2 + \tilde{M} V_k)\delta^2\right)}{(2G^2 + \tilde{M} V_k)^2} \\ B_k &= \frac{(2G^2 V_k + (\tilde{M} - m^2) V_k^2)((2G^2 + \tilde{M} V_k)(1 - \delta^2) - c^2)}{(2G^2 + \tilde{M} V_k)^2} \\ Z_k &= \frac{\left(2G^2 + \tilde{M} V_k\right)(\delta^2 + \zeta_k) + c^2}{(2G^2 + \tilde{M} V_k)(1 - \delta^2) - c^2} \end{aligned}$$

Note that $A_k \geq 0$ and $B_k \geq 0$ due to the condition $2G^2(1 - \delta^2) \geq c^2$. The objective in (61) therefore has a form very similar to the objective in (26). Applying Lemma 1, we deduce that $V_{k+1} = (\sqrt{A_k} + \sqrt{B_k})^2$, which is the same as (49). The associated Z_k^{opt} is $\sqrt{\frac{A_k}{B_k}}$. To ensure this is a feasible choice, it remains to check that the associated $\zeta_k^{\text{opt}} > 0$ as well. Via algebraic manipulations, one can show that $\zeta_k > 0$ is equivalent to $V_k > V_\star$. We can also verify A_k is a monotonically increasing function of V_k, and B_k is a monotonically nondecreasing function of V_k. Hence h is a monotonically increasing function. Also notice $V_\star$ is a fixed point of (49). Therefore, if we assume $V_k = h(V_{k-1})$ and $V_{k-1} > V_\star$, we have $V_k = h(V_{k-1}) > h(V_\star) = V_\star$. Hence we guarantee $\zeta_k > 0$

and $V_{k+1} = h(V_k)$. By similar arguments, one can verify $V_1 = h(V_0)$. And it is assumed that $V_0 > V_\star$. This completes the induction argument.

Item 2 follows from a similar argument to the one used in Sect. 2.2. Finally, Item 3 can be proven by choosing a sufficiently small constant stepsize α to make $\hat{U}_k$ arbitrarily close to $V_\star$. Since $V_\star \leq V_k \leq \hat{U}_k$, we conclude that $\lim_{k\to\infty} V_k = V_\star$, as required. □

References

1. Agarwal, A., Bartlett, P.L., Ravikumar, P., Wainwright, M.J.: Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. IEEE Trans. Inf. Theory **58**(5), 3235–3249 (2012)
2. Arora, S., Ge, R., Ma, T., Moitra, A.: Simple, efficient, and neural algorithms for sparse coding. In: Conference on Learning Theory, pp. 113–149 (2015)
3. Bertsekas, D.: Nonlinear Programming, 2nd edn. Athena scientific, Belmont (2002)
4. Bottou, L.: Large-scale machine learning with stochastic gradient descent. In: Proceedings of COMPSTAT'2010, pp. 177–186 (2010)
5. Bottou, L., Curtis, F., Nocedal, J.: Optimization methods for large-scale machine learning. SIAM Rev. **60**(2), 223–311 (2018)
6. Bottou, L., LeCun, Y.: Large scale online learning. Adv. Neural Inf. Process. Syst. **16**, 217 (2004)
7. Bubeck, S.: Convex optimization: algorithms and complexity. Found. Trends® Mach. Learn. **8**(3–4), 231–357 (2015)
8. Chen, Y., Candes, E.: Solving random quadratic systems of equations is nearly as easy as solving linear systems. In: Advances in Neural Information Processing Systems, pp. 739–747 (2015)
9. d'Aspremont, A.: Smooth optimization with approximate gradient. SIAM J. Optim. **19**(3), 1171–1183 (2008)
10. De Klerk, E., Glineur, F., Taylor, A.: On the worst-case complexity of the gradient method with exact line search for smooth strongly convex functions. Optim. Lett. **11**(7), 1185–1199 (2017)
11. Defazio, A., Bach, F., Lacoste-Julien, S.: Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In: Advances in Neural Information Processing Systems (2014)
12. Defazio, A., Domke, J., Caetano, T.: Finito: A faster, permutable incremental gradient method for big data problems. In: Proceedings of the 31st International Conference on Machine Learning, pp. 1125–1133 (2014)
13. Devolder, O., Glineur, F., Nesterov, Y.: First-order methods of smooth convex optimization with inexact oracle. Math. Program. **146**(1–2), 37–75 (2014)
14. Drori, Y., Teboulle, M.: Performance of first-order methods for smooth convex minimization: a novel approach. Math. Program. **145**(1–2), 451–482 (2014)
15. Feyzmahdavian, H., Aytekin, A., Johansson, M.: A delayed proximal gradient method with linear convergence rate. In: 2014 IEEE International Workshop on Machine Learning for Signal Processing, pp. 1–6 (2014)
16. Grant, M., Boyd, S.: Graph implementations for nonsmooth convex programs. In: Blondel, V., Boyd, S., Kimura, H. (eds.) Recent Advances in Learning and Control. Lecture Notes in Control and Information Sciences, pp. 95–110. Springer (2008). http://stanford.edu/~boyd/graph_dcp.html
17. Grant, M., Boyd, S.: CVX: Matlab software for disciplined convex programming, version 2.1. http://cvxr.com/cvx (2014)
18. Hu, B., Seiler, P., Rantzer, A.: A unified analysis of stochastic optimization methods using jump system theory and quadratic constraints. In: Proceedings of the 2017 Conference on Learning Theory, vol. 65, pp. 1157–1189 (2017)
19. Johnson, R., Zhang, T.: Accelerating stochastic gradient descent using predictive variance reduction. In: Advances in Neural Information Processing Systems, pp. 315–323 (2013)
20. Lee, J.C., Valiant, P.: Optimizing star-convex functions. In: 2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS), pp. 603–614 (2016)

21. Lessard, L., Recht, B., Packard, A.: Analysis and design of optimization algorithms via integral quadratic constraints. SIAM J. Optim. **26**(1), 57–95 (2016)
22. Moulines, E., Bach, F.: Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In: Advances in Neural Information Processing Systems, pp. 451–459 (2011)
23. Nedić, A., Bertsekas, D.: Convergence rate of incremental subgradient algorithms. In: Stochastic Optimization: Algorithms and Applications, pp. 223–264 (2001)
24. Needell, D., Ward, R., Srebro, N.: Stochastic gradient descent, weighted sampling, and the randomized Kaczmarz algorithm. In: Advances in Neural Information Processing Systems, pp. 1017–1025 (2014)
25. Nishihara, R., Lessard, L., Recht, B., Packard, A., Jordan, M.: A general analysis of the convergence of ADMM. In: Proceedings of the 32nd International Conference on Machine Learning, pp. 343–352 (2015)
26. Robbins, H., Monro, S.: A stochastic approximation method. Ann. Math. Stat. **22**(3), 400–407 (1951)
27. Roux, N., Schmidt, M., Bach, F.: A stochastic gradient method with an exponential convergence rate for strongly-convex optimization with finite training sets. In: Advances in Neural Information Processing Systems (2012)
28. Schmidt, M., Roux, N., Bach, F.: Minimizing finite sums with the stochastic average gradient. Math. Program. **162**(1–2), 83–112 (2017)
29. Schmidt, M., Roux, N.L., Bach, F.R.: Convergence rates of inexact proximal-gradient methods for convex optimization. In: Advances in Neural Information Processing Systems, pp. 1458–1466 (2011)
30. Shalev-Shwartz, S., Zhang, T.: Stochastic dual coordinate ascent methods for regularized loss. J. Mach. Learn. Res. **14**(1), 567–599 (2013)
31. Sun, R., Luo, Z.Q.: Guaranteed matrix completion via non-convex factorization. IEEE Trans. Inf. Theory **62**(11), 6535–6579 (2016)
32. Taylor, A., Bach, F.: Stochastic first-order methods: non-asymptotic and computer-aided analyses via potential functions. In: Proceedings of the 2019 Conference on Learning Theory, pp. 2934–2992 (2019)
33. Taylor, A., Hendrickx, J., Glineur, F.: Smooth strongly convex interpolation and exact worst-case performance of first-order methods. Math. Program. **161**(1–2), 307–345 (2017)
34. Taylor, A., Hendrickx, J.M., Glineur, F.: Exact worst-case performance of first-order methods for composite convex optimization. SIAM J. Optim. **27**(3), 1283–1313 (2017)
35. Taylor, A., Van Scoy, B., Lessard, L.: Lyapunov functions for first-order methods: Tight automated convergence guarantees. In: Proceedings of the 35th International Conference on Machine Learning, pp. 4897–4906 (2018)
36. Teo, C., Smola, A., Vishwanathan, S., Le, Q.: A scalable modular convex solver for regularized risk minimization. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 727–736 (2007)