Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Technical Note—On the Convergence Rate of Stochastic Approximation for Gradient-Based Stochastic Optimization

Jiaqiao Hu, Michael C. Fu

To cite this article:
Jiaqiao Hu, Michael C. Fu (2024) Technical Note—On the Convergence Rate of Stochastic Approximation for Gradient-Based Stochastic Optimization. Operations Research

Published online in Articles in Advance 08 Mar 2024

. https://doi.org/10.1287/opre.2023.0055

https://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Articles in Advance, pp. 1–8
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Methods

Technical Note—On the Convergence Rate of Stochastic Approximation for Gradient-Based Stochastic Optimization

Jiaqiao Hu,[a] Michael C. Fu[b,*]

[a] Department of Applied Mathematics and Statistics, State University of New York at Stony Brook, Stony Brook, New York 11794; [b] Robert H. Smith School of Business & Institute for Systems Research, University of Maryland, College Park, Maryland 20742

*Corresponding author

Contact: jqhu@ams.sunysb.edu, https://orcid.org/0000-0002-9999-672X (JH); mfu@umd.edu, https://orcid.org/0000-0003-2105-4932 (MCF)

Received: February 1, 2023
Revised: September 8, 2023
Accepted: January 11, 2024
Published Online in Articles in Advance: March 8, 2024

Area of Review: Simulation

https://doi.org/10.1287/opre.2023.0055

Abstract. We consider stochastic optimization via gradient-based search. Under a stochastic approximation framework, we apply a recently developed convergence rate analysis to provide a new finite-time error bound for a class of problems with convex differentiable structures. For noisy black-box functions, our main result allows us to derive finite-time bounds in the setting where the gradients are estimated via finite-difference estimators, including those based on randomized directions such as the simultaneous perturbation stochastic approximation algorithm. In particular, the convergence rate analysis sheds light on when it may be advantageous to use such randomized gradient estimates in terms of problem dimension and noise levels.

Funding: This work was supported by the Air Force Office of Scientific Research [Grant FA95502010211] and the National Science Foundation [Grants IIS-2123684 and CMMI-2027527].

Keywords: stochastic approximation • convergence rate • finite-time analysis • finite differences • random directions • simultaneous perturbation

1. Introduction

Gradient-based algorithms are the most commonly used methods for addressing continuous optimization problems with some known (or assumed) smoothness. Focusing on the stochastic setting where the gradient is estimated with both bias (e.g., finite differences) and noise, in this note we use the finite-time analysis introduced in Hu et al. (2024) to derive new error bounds on the iterates in gradient-based search for a class of problems with convex differentiable structures. Our analysis focuses on the standard stochastic approximation (SA) algorithm with diminishing step-sizes, which complements recent developments in the machine learning literature (e.g., Duchi et al. 2015, Karimi et al. 2019, Driggs et al. 2022, Demidovich et al. 2023) that consider constant step-sizes and/or variants of such algorithms. In addition, as contrasted with existing studies, which commonly assume a bounded gradient estimation error at the optimum (see, e.g., Duchi et al. 2015, Bottou et al. 2018, Chen and Luss 2019, Hu et al. 2021, Demidovich et al. 2023, and references therein), our results are based on an explicit bias-variance decomposition, where the variance of the gradient estimator is allowed to increase with the number of algorithm iterations. Such a scenario frequently arises in traditional stochastic approximation settings (Kushner and Yin 1997), for example, when a finite-difference (FD)-based estimator is constructed based on a sequence of diminishing perturbation sizes (Kiefer and Wolfowitz 1952, Spall 1992).

In the noisy black-box setting where only noisy evaluations of the output function are available, the main theoretical result is used to compare the finite-time performance of traditional Kiefer and Wolfowitz (1952) (KW) algorithms and randomized finite-difference gradient-based search, such as simultaneous perturbation stochastic approximation (SPSA) of Spall (1992). Specifically, the convergence rate analysis enables the characterization of finite-time performance of finite-difference-based gradient search in terms of problem dimension and noise levels, providing guidance on when it might be appropriate to use randomized gradients and bridging a gap between the asymptotic analysis of Spall (1992) and Kushner and Yin (1997) and the "static" gradient bias-variance trade-off analysis of Scheinberg (2022) and Berahas et al. (2022).

The remainder of this note is organized as follows. Section 2 presents the optimization problem setting. The stochastic approximation framework and main result are presented in Section 3, and Section 4 specializes to the black-box setting where finite-difference-based estimates are used for the gradient. Section 5 concludes.

2. Problem Setting

Consider the following optimization problem:

$$\min_{\boldsymbol{\theta}\in\Theta} F(\boldsymbol{\theta}), \tag{1}$$

where the feasible region $\Theta \subseteq \Re^d$ is a convex set that is either compact or the entire (unbounded) space, that is, $\Theta = \Re^d$. We assume that the objective function $F : \Re^d \to \Re$ is smooth and strongly convex, satisfying the regularity properties stated below.

Assumption A1 (Differentiability). *The function $F(\boldsymbol{\theta})$ is twice continuously differentiable on Θ.*

Assumption A2 (Lipschitz Smoothness). *There exists a constant $M > 0$ such that $\|\nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_1} - \nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_2}\| \le M\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|$ for all $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$, where $\|\cdot\|$ is the Euclidean norm.*

Assumption A3 (Strong Convexity). *Let $\lambda(\boldsymbol{\theta})$ be the smallest eigenvalue of the Hessian matrix $\boldsymbol{H}(\boldsymbol{\theta}) := \nabla^2_{\boldsymbol{\theta}} F(\boldsymbol{\theta})$. There exists a constant $m > 0$ such that $\lambda(\boldsymbol{\theta}) \ge m$ for all $\boldsymbol{\theta} \in \Theta$.*

Assumptions A1–A3 are standard conditions frequently adopted when analyzing the convergence rates of gradient descent methods (e.g., Ghadimi and Lan 2012, Bottou et al. 2018, Berahas et al. 2022, Scheinberg 2022). Note that when Θ is compact, Assumption A2 follows automatically from Assumption A1. Moreover, the constant M in Assumption A2 also serves as an upper bound on the largest eigenvalue of $\boldsymbol{H}(\boldsymbol{\theta})$. We focus on the setting where F is a black-box function estimated with noise. For example, in a stochastic optimization setting, the objective function is often given in the form of an expectation, where the value of $F(\boldsymbol{\theta})$ may not be computed analytically but in general can be estimated by sampling from an underlying stochastic/simulation model.

3. Convergence Rate Analysis for General Gradient-Based Search

Under the assumptions stated in Section 2, Problem (1) has a unique solution $\boldsymbol{\theta}^*$, which can be found through gradient-based search taking the following general form, known as stochastic approximation:

$$\boldsymbol{\theta}_{k+1} = \Pi_\Theta(\boldsymbol{\theta}_k - a_k \boldsymbol{g}_k(\boldsymbol{\theta}_k)), \tag{2}$$

where $\boldsymbol{g}_k(\boldsymbol{\theta}_k)$ is an estimate of the true gradient $\boldsymbol{G}(\boldsymbol{\theta}_k) := \nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_k}$ obtained at iteration k, $a_k > 0$ is the step-size, and $\Pi_\Theta(\cdot)$ is the projection operator that brings an iterate of (2) back into the feasible region Θ whenever it becomes infeasible. For the unbounded case $\Theta = \Re^d$, the projection operator should be removed.

Let $\mathcal{F}_k = \sigma\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k\}$ be an increasing σ-field generated by the sequence of random iterates $\{\boldsymbol{\theta}_k\}$ obtained up to iteration k. The gradient estimation error $\boldsymbol{g}_k(\boldsymbol{\theta}_k) - \boldsymbol{G}(\boldsymbol{\theta}_k)$ can then be written as the sum of a bias $\boldsymbol{b}_k := E[\boldsymbol{g}_k(\boldsymbol{\theta}_k)|\mathcal{F}_k] - \boldsymbol{G}(\boldsymbol{\theta}_k)$ and a noise term $\boldsymbol{\epsilon}_k := \boldsymbol{g}_k(\boldsymbol{\theta}_k) - E[\boldsymbol{g}_k(\boldsymbol{\theta}_k)|\mathcal{F}_k]$. Under fairly mild conditions on a_k, $\boldsymbol{b}_k$, and $\boldsymbol{\epsilon}_k$, the sequence $\{\boldsymbol{\theta}_k\}$ can be shown to converge to $\boldsymbol{\theta}^*$ with probability one (e.g., Kushner and Yin 1997). Our goal is to characterize the finite-time convergence rate of (2) in terms of the gradient estimation bias $\boldsymbol{b}_k$ and noise $\boldsymbol{\epsilon}_k$. Throughout the analysis, for two positive real sequences $\{v_k\}$ and $\{w_k\}$, we write $v_k = o(w_k)$ if $\lim_{k\to\infty} v_k/w_k = 0$; $v_k = O(w_k)$ if $\exists C > 0$ and $\exists K > 0$ such that $v_k \le C w_k$ for all $k \ge K$. We consider standard step-size sequences $\{a_k\}$ of the form $a_k = a/k^\alpha$ for constants $a > 0$, $\alpha \in [0, 1)$, where the case $\alpha = 0$ corresponds to constant step-sizes and otherwise diminishing step-sizes, and impose the following conditions on $\boldsymbol{\theta}^*$ and the gradient estimation error:

Assumption A4. *The optimal solution $\boldsymbol{\theta}^*$ is an interior point of Θ.*

Assumption A5. *There are constants $q_1 > 0$, $q_2 < \alpha/2$, and $C_1, C_2 > 0$ such that $E[\|\boldsymbol{b}_k\|^2] \le C_1 k^{-2q_1}$ and $E[\|\boldsymbol{\epsilon}_k\|^2] \le C_2 k^{2q_2}$.*

Assumption A4 holds trivially for unconstrained problems. Assumption A5 requires the estimation bias to be asymptotically negligible, whereas the noise variance only has to go to zero when scaled by the step-size a_k, that is, $a_k E[\|\boldsymbol{\epsilon}_k\|^2] \to 0$ because $2q_2 < \alpha$. Other than the order of the gradient $\boldsymbol{g}_k$ noise variance, we do not impose any additional assumptions on the measurement noise of the underlying function F. For example, the measurement noise could be i.i.d., martingale differences, or correlated in general. Note that the magnitude of the constants q_1 and q_2 can be arbitrarily large. The $q_1 = \infty$ case means that there is no estimation bias, whereas $q_2 = -\infty$ corresponds to deterministic gradient estimation.

Assumption A5 differs from most assumptions used in existing studies on biased stochastic gradient descent (see, e.g., section 4 of Demidovich et al. 2023), which are based on bounding the moments of $\boldsymbol{g}_k$ with those of the true gradient $\boldsymbol{G}$. This implies that the gradient estimation error will either vanish or remain uniformly bounded at the optimum, which is generally not satisfied for settings where a finite-difference gradient estimator with a sequence of diminishing perturbation sizes is employed. We now state and prove our main result, which provides a finite-time bound on the mean absolute error of the gradient descent method (2).

Theorem 1 (Diminishing Step-Sizes). *Let* Assumptions A1–A5 *hold and $\rho < m$ be a positive constant. For $a_k = a/k^\alpha$, $\alpha \in (0, 1)$, $a > 0$, if a_k is chosen such that $a_k < \min\{2(m-\rho)/M^2, 1/(2\rho)\}$ and $(1 + Ma_k)/\sqrt{1 - 2\rho a_k} \le 2$ $\forall k \ge 0$, then the sequence $\{\boldsymbol{\theta}_k\}$ generated by (2) satisfies*

$$E[\|\boldsymbol{\theta}_k - \boldsymbol{\theta}^*\|] \le Ce^{-\rho\sum_{i=1}^{k-1}\alpha_i} + \frac{2\sqrt{C_1}}{\rho}k^{-q_1} + \frac{\sqrt{C_2}}{\sqrt{\rho}}a_k^{\frac{1}{2}}k^{q_2} + o(k^{-q_1}) + o(a_k^{\frac{1}{2}}k^{q_2}) \tag{3}$$

for some constant $\mathcal{C} > 0$. In addition,

$$\sqrt{E[F(\boldsymbol{\theta}_k) - F(\boldsymbol{\theta}^*)]} \le \bar{\mathcal{C}} e^{-\rho \sum_{i=1}^{k-1} \alpha_i} + \sqrt{\frac{M}{2}} \left(\frac{2\sqrt{C_1}}{\rho} k^{-q_1} + \frac{\sqrt{C_2}}{\sqrt{\rho}} a_k^{\frac{1}{2}} k^{q_2} \right) + o(k^{-q_1}) + o(a_k^{\frac{1}{2}} k^{q_2})$$

for some constant $\bar{\mathcal{C}} > 0$.

Proof. Note that Assumption A4 implies that the projection operation in (2) will not have an effect on the convergence rate of the algorithm (e.g., Kushner and Yin 1997, Hu et al. 2024). The operator Π_Θ will henceforth be dropped in our analysis without loss of generality.

Let $\boldsymbol{\eta}_k := \boldsymbol{\theta}_k - \boldsymbol{\theta}^*$. We can write (2) in terms of $\boldsymbol{\eta}_k$ as $\boldsymbol{\eta}_{k+1} = \boldsymbol{\eta}_k - a_k \boldsymbol{G}(\boldsymbol{\theta}_k) - a_k(\boldsymbol{g}_k(\boldsymbol{\theta}_k) - \boldsymbol{G}(\boldsymbol{\theta}_k))$. It follows that

$$\begin{aligned} \|\boldsymbol{\eta}_{k+1}\|^2 = \|\boldsymbol{\eta}_k\|^2 + a_k^2 \|\boldsymbol{G}(\boldsymbol{\theta}_k)\|^2 + a_k^2 \|\boldsymbol{g}_k(\boldsymbol{\theta}_k) - \boldsymbol{G}(\boldsymbol{\theta}_k)\|^2 \\ - 2a_k \boldsymbol{\eta}_k^T \boldsymbol{G}(\boldsymbol{\theta}_k) - 2a_k \boldsymbol{\eta}_k^T (\boldsymbol{g}_k(\boldsymbol{\theta}_k) - \boldsymbol{G}(\boldsymbol{\theta}_k)) \\ + 2a_k^2 \boldsymbol{G}^T(\boldsymbol{\theta}_k)(\boldsymbol{g}_k(\boldsymbol{\theta}_k) - \boldsymbol{G}(\boldsymbol{\theta}_k)). \end{aligned}$$

Because $\boldsymbol{G}(\boldsymbol{\theta}^*) = 0$, by Assumption A1 and the mean value theorem, $\boldsymbol{G}(\boldsymbol{\theta}_k) = \boldsymbol{G}(\boldsymbol{\theta}^*) + \boldsymbol{H}(\overline{\boldsymbol{\theta}}_k)(\boldsymbol{\theta}_k - \boldsymbol{\theta}^*) = \boldsymbol{H}(\overline{\boldsymbol{\theta}}_k)\boldsymbol{\eta}_k$ for some $\overline{\boldsymbol{\theta}}_k$ on the line segment between $\boldsymbol{\theta}_k$ and $\boldsymbol{\theta}^*$. Under Assumptions A2 and A3, the Rayleigh-Ritz inequality (e.g., Rugh 1996) implies that $\boldsymbol{\eta}_k^T \boldsymbol{G}(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k^T \boldsymbol{H}(\overline{\boldsymbol{\theta}}_k)\boldsymbol{\eta}_k \ge m\|\boldsymbol{\eta}_k\|^2$ and $\|\boldsymbol{G}(\boldsymbol{\theta}_k)\|^2 = \boldsymbol{\eta}_k^T \boldsymbol{H}^T(\overline{\boldsymbol{\theta}}_k)\boldsymbol{H}(\overline{\boldsymbol{\theta}}_k)\boldsymbol{\eta}_k \le M^2\|\boldsymbol{\eta}_k\|^2$, where recall that by Assumption A2, M is an upper bound on the largest eigenvalue of $\boldsymbol{H}(\boldsymbol{\theta})$. Using these bounds, we get that

$$\begin{aligned} \|\boldsymbol{\eta}_{k+1}\|^2 \le (1 - 2ma_k + M^2 a_k^2)\|\boldsymbol{\eta}_k\|^2 + a_k^2 \|\boldsymbol{g}_k(\boldsymbol{\theta}_k) - \boldsymbol{G}(\boldsymbol{\theta}_k)\|^2 \\ -2a_k \boldsymbol{\eta}_k^T (\boldsymbol{g}_k(\boldsymbol{\theta}_k) - \boldsymbol{G}(\boldsymbol{\theta}_k)) + 2a_k^2 \boldsymbol{G}^T(\boldsymbol{\theta}_k)(\boldsymbol{g}_k(\boldsymbol{\theta}_k) - \boldsymbol{G}(\boldsymbol{\theta}_k)). \end{aligned}$$

Taking conditional expectations on both sides and using the fact $E[\|\boldsymbol{g}_k(\boldsymbol{\theta}_k) - \boldsymbol{G}(\boldsymbol{\theta}_k)\|^2 | \mathcal{F}_k] = E[\|\boldsymbol{b}_k\|^2 | \mathcal{F}_k] + E[\|\boldsymbol{\epsilon}_k\|^2 | \mathcal{F}_k]$, then yield

$$\begin{aligned} E[\|\boldsymbol{\eta}_{k+1}\|^2 | \mathcal{F}_k] \le (1 - 2ma_k + M^2 a_k^2)\|\boldsymbol{\eta}_k\|^2 \\ + a_k^2 E[\|\boldsymbol{b}_k\|^2 | \mathcal{F}_k] + a_k^2 E[\|\boldsymbol{\epsilon}_k\|^2 | \mathcal{F}_k] \\ - 2a_k \boldsymbol{\eta}_k^T \boldsymbol{b}_k + 2a_k^2 \boldsymbol{G}^T(\boldsymbol{\theta}_k)\boldsymbol{b}_k. \end{aligned}$$

By unconditioning on $\mathcal{F}_k$, applying the Cauchy-Schwarz inequality and the conditions $a_k < \min\{2(m-\rho)/M^2, 1/(2\rho)\}$, $(1 + Ma_k)/\sqrt{1 - 2\rho a_k} \le 2$ for all k, we further obtain

$$\begin{aligned} E[\|\boldsymbol{\eta}_{k+1}\|^2] &\le (1 - 2ma_k + M^2 a_k^2) E[\|\boldsymbol{\eta}_k\|^2] \\ &\quad + 2a_k(1 + Ma_k)\sqrt{E[\|\boldsymbol{\eta}_k\|^2]}\sqrt{E[\|\boldsymbol{b}_k\|^2]} \\ &\quad + a_k^2 E[\|\boldsymbol{b}_k\|^2] + a_k^2 E[\|\boldsymbol{\epsilon}_k\|^2] \\ &\le (1 - 2\rho a_k) E[\|\boldsymbol{\eta}_k\|^2] + 2a_k(1 + Ma_k)\sqrt{E[\|\boldsymbol{\eta}_k\|^2]} \\ &\quad \sqrt{E[\|\boldsymbol{b}_k\|^2]} + a_k^2 E[\|\boldsymbol{b}_k\|^2] + a_k^2 E[\|\boldsymbol{\epsilon}_k\|^2] \\ &\le \left(\sqrt{1 - 2\rho a_k}\sqrt{E[\|\boldsymbol{\eta}_k\|^2]} + a_k \frac{(1 + Ma_k)}{\sqrt{1 - 2\rho a_k}} \sqrt{E[\|\boldsymbol{b}_k\|^2]} \right)^2 \\ &\quad + a_k^2 E[\|\boldsymbol{\epsilon}_k\|^2] \\ &\le ((1 - \rho a_k)\sqrt{E[\|\boldsymbol{\eta}_k\|^2]} + 2a_k \sqrt{E[\|\boldsymbol{b}_k\|^2]})^2 + a_k^2 E[\|\boldsymbol{\epsilon}_k\|^2] \\ &\le ((1 - \rho a_k)\sqrt{E[\|\boldsymbol{\eta}_k\|^2]} + 2a_k \sqrt{C_1} k^{-q_1})^2 + a_k^2 C_2 k^{2q_2}, \end{aligned}$$

where in the penultimate step we have used the inequality $\sqrt{1 - x} \le 1 - x/2$ for $x \in [0, 1]$, and the last inequality follows from Assumption A5.

Now consider the sequence of mappings $\mathcal{T}_k(\cdot)$, $k = 1, 2, \ldots$, defined by

$$\mathcal{T}_k(x) = \sqrt{((1 - \rho a_k)x + 2a_k \sqrt{C_1} k^{-q_1})^2 + a_k^2 C_2 k^{2q_2}}.$$

It can be readily verified that for each k, $\mathcal{T}_k$ is a contraction mapping with $|\mathcal{T}_k(x) - \mathcal{T}_k(y)| \le (1 - \rho a_k)|x - y|$, and its unique fixed point x_k^* satisfies $x_k^* \le 2(\sqrt{C_1}/\rho) k^{-q_1} + (\sqrt{C_2}/\sqrt{\rho})\sqrt{a_k} k^{q_2}$.

Next, let $x_1 = \sqrt{E[\|\boldsymbol{\eta}_1\|^2]}$ and consider the sequence $\{x_k\}$ generated by $x_{k+1} = \mathcal{T}_k(x_k)$, $k = 1, 2, \ldots$. By induction, we have $\sqrt{E[\|\boldsymbol{\eta}_k\|^2]} \le x_k$ for all k. On the other hand, by the contraction property of $\mathcal{T}_k$ and repeated application of the triangle inequality,

$$\begin{aligned} |x_{k+1} - x_{k+1}^*| &\le |x_{k+1} - x_k^*| + |x_k^* - x_{k+1}^*| \\ &= |\mathcal{T}_k(x_k) - \mathcal{T}_k(x_k^*)| + |x_k^* - x_{k+1}^*| \\ &\le (1 - a_k\rho)|x_k - x_k^*| + |x_k^* - x_{k+1}^*| \\ &\le (1 - a_k\rho)|x_k - x_{k-1}^*| \\ &\quad + (1 - a_k\rho)|x_k^* - x_{k-1}^*| + |x_k^* - x_{k+1}^*| \\ &\cdots \\ &\le \prod_{i=1}^{k} (1 - \alpha_i\rho)|x_1 - x_1^*| \\ &\quad + \sum_{i=1}^{k} \left[\prod_{j=i+1}^{k} (1 - \alpha_j\rho) \right] \alpha_i\rho \frac{|x_{i+1}^* - x_i^*|}{\alpha_i\rho}. \end{aligned} \tag{4}$$

To finish the proof of the first part, we need the following intermediate result, which is a strengthened version of lemma 3 in Hu et al. (2024) and whose proof is given in the appendix.

Lemma 1. *Let $u(i) = s/i^p$ and $w(i) = O(1/i^q)$, where $s > 0$, $p \in (0,1)$, $q > 0$, and $w(i) > 0$ for all $i = 1, 2, \ldots$. Then*

$$\sum_{i=1}^{k}\left[\prod_{j=i+1}^{k}(1-u(j))\right]u(i)w(i) = O(k^{-q}).$$

For the specific form of the step-size $a_k = a/k^\alpha$, it can be shown that $|x_k^* - x_{k+1}^*| = O(k^{-q_1-1}) + O(k^{-\alpha/2+q_2-1})$. Using the fact that $\prod_{i=1}^{k}(1-\alpha_i\rho) \le \exp(-\rho\sum_{i=1}^{k}\alpha_i)$ and applying Lemma 1 to the second term on the right-hand side of (4),

$$|x_{k+1} - x_{k+1}^*| \le \mathcal{C}e^{-\rho\sum_{i=1}^{k}\alpha_i} + o(k^{-q_1}) + o(a_k^{\frac{1}{2}}k^{q_2})$$

for some constant $\mathcal{C}$. Finally, the first result is proved by noticing that $E[\|\boldsymbol{\theta}_k - \boldsymbol{\theta}^*\|] \le \sqrt{E[\|\boldsymbol{\eta}_k\|^2]} \le x_k \le x_k^* + \mathcal{C}e^{-\rho\sum_{i=1}^{k-1}\alpha_i} + o(k^{-q_1}) + o(\sqrt{a_k}k^{q_2})$.

To show the second part of the theorem, we note that by the Lipschitz smoothness condition A2 and the fact that $\boldsymbol{G}(\boldsymbol{\theta}^*) = 0$,

$$\begin{aligned} F(\boldsymbol{\theta}_k) &\le F(\boldsymbol{\theta}^*) + \boldsymbol{G}^T(\boldsymbol{\theta}^*)(\boldsymbol{\theta}_k - \boldsymbol{\theta}^*) + \frac{M}{2}\|\boldsymbol{\theta}_k - \boldsymbol{\theta}^*\|^2 \\ &= F(\boldsymbol{\theta}^*) + \frac{M}{2}\|\boldsymbol{\theta}_k - \boldsymbol{\theta}^*\|^2. \end{aligned}$$

Hence, we obtain $\sqrt{E[F(\boldsymbol{\theta}_k) - F(\boldsymbol{\theta}^*)]} \le \sqrt{M/2}\sqrt{E[\|\boldsymbol{\eta}_k\|^2]}$, and the desired result immediately follows as a consequence of the first part of the theorem. □

Under Assumption A5, because $\sqrt{E[\|\boldsymbol{b}_k\|^2]} \le \sqrt{C_1}k^{-q_1}$ and $\sqrt{E[\|\boldsymbol{\epsilon}_k\|^2]} \le \sqrt{C_2}k^{q_2}$, Theorem 1 essentially gives a performance bound for (2) in terms of the bounds on the gradient estimation bias and variance. When the orders of the errors $E[\|\boldsymbol{b}_k\|^2]$ and $E[\|\boldsymbol{\epsilon}_k\|^2]$ are uniform in k, the result can be stated as $E[\|\boldsymbol{\theta}_k - \boldsymbol{\theta}^*\|] = O(e^{-\rho\sum_{i=1}^{k-1}\alpha_i}) + O(\sqrt{E[\|\boldsymbol{b}_k\|^2]}) + O(a_k^{1/2}\sqrt{E[\|\boldsymbol{\epsilon}_k\|^2]})$. This seems to be the strongest possible result in this generality and conforms to existing work on the (asymptotic) convergence rate of SA algorithms. Supposing that the estimator $\boldsymbol{g}_k(\boldsymbol{\theta}_k)$ can be obtained without bias and with a constant variance (i.e., $q_1 = \infty$ and $q_2 = 0$), then the performance bound diminishes at the rate $O(a_k^{1/2})$. On the other hand, if there is no estimation noise (i.e., $q_2 = -\infty$), then the rate is primarily dominated by the order of the estimation bias, and the algorithm converges geometrically when the exact gradient is used. Another observation from Theorem 1 is that the bound in (3) is decreasing in ρ and hence can be made small by taking the value of ρ close to m. Because m is a lower bound on the smallest eigenvalue of the Hessian of F, which roughly measures the degree of convexity of the function, faster convergence rates can generally be expected for objective functions with stronger convex curvatures.

When constant step-sizes are used (i.e., $\alpha = 0$), and also assuming constant gradient estimation bias and variance, a result analogous to Theorem 1 is obtained below.

Corollary 1 (Constant Step-Sizes). *Let Assumptions A1–A4 hold. Suppose that the conditions on a_k in Theorem 1 hold with $a_k = a$ for all k and that Assumption A5 is satisfied with $q_1 = q_2 = 0$, that is, $E[\|\boldsymbol{b}_k\|^2] \le C_1$ and $E[\|\boldsymbol{\epsilon}_k\|^2] \le C_2$. Then the sequence $\{\boldsymbol{\theta}_k\}$ generated by (2) satisfies*

$$E[\|\boldsymbol{\theta}_k - \boldsymbol{\theta}^*\|] \le \mathcal{C}(1-a\rho)^{k-1} + \frac{2\sqrt{C_1}}{\rho} + \frac{\sqrt{C_2}}{\sqrt{\rho}}a^{\frac{1}{2}}$$

for some constant $\mathcal{C} > 0$. In addition,

$$\sqrt{E[F(\boldsymbol{\theta}_k) - F(\boldsymbol{\theta}^*)]} \le \bar{\mathcal{C}}(1-a\rho)^{k-1} + \sqrt{\frac{M}{2}}\left(\frac{2\sqrt{C_1}}{\rho} + \frac{\sqrt{C_2}}{\sqrt{\rho}}a^{\frac{1}{2}}\right)$$

for some constant $\bar{\mathcal{C}} > 0$.

Proof. The proof is almost identical to that of Theorem 1 with a replacing a_k and $q_1 = q_2 = 0$. The only difference is that the contraction mapping $\mathcal{T}_k$ defined in Theorem 1 no longer depends on k and reduces to

$$\mathcal{T}(x) = \sqrt{((1-a\rho)x + 2a\sqrt{C_1})^2 + a^2C_2},$$

with fixed point satisfying $x^* \le 2(\sqrt{C_1}/\rho) + (\sqrt{C_2}/\sqrt{\rho})\sqrt{a}$. Again, by defining $\boldsymbol{\eta}_k = \boldsymbol{\theta}_k - \boldsymbol{\theta}^*$ and letting $x_1 = \sqrt{E[\|\boldsymbol{\eta}_1\|^2]}$ and $\{x_k\}$ be generated by $x_{k+1} = \mathcal{T}(x_k)$, $k = 1, 2, \ldots$, we have

$$\begin{aligned} |x_{k+1} - x^*| = |\mathcal{T}(x_k) - \mathcal{T}(x^*)| &\le (1-a\rho)|x_k - x^*| \\ &\le \cdots \le (1-a\rho)^k|x_1 - x^*|. \end{aligned}$$

Consequently, the first result is proved by noticing that $\sqrt{E[\|\boldsymbol{\eta}_k\|^2]} \le x_k \le x^* + \mathcal{C}(1-a\rho)^{k-1}$ for some constant $\mathcal{C}$. The proof of the second result follows the analogous proof in Theorem 1. □

4. Finite-Difference-Based Gradient Descent Algorithms

We now specialize our analysis to specific gradient descent algorithms and discuss their relative advantages/disadvantages based on the performance bound stated in Theorem 1. We consider central/symmetric finite-difference gradient approximation schemes. Because their analysis typically relies on a third-order Taylor series expansion of the objective function, condition A1 will thus be strengthened to require $F(\boldsymbol{\theta})$ to be three-times continuously differentiable. Throughout this section, we let $\nabla^3_{\boldsymbol{\theta}}F(\boldsymbol{\theta})$ be the tensor of F and assume that its elements are uniformly bounded on an open neighborhood $\mathcal{N}(\Theta)$ of Θ. Let $L := \sup_{\boldsymbol{\theta}\in\mathcal{N}(\Theta)}\max_{1\le i,j,k\le d}|[\nabla^3_{\boldsymbol{\theta}}F(\boldsymbol{\theta})]_{i,j,k}|$. We also define $L_G := \sup_{\boldsymbol{\theta}\in\Theta}\max_{1\le i\le d}|[\boldsymbol{G}(\boldsymbol{\theta})]_i|$ so that a conservative bound on $\|\boldsymbol{G}(\boldsymbol{\theta}_k)\|^2$ is given by dL_G^2. The perturbation-

sizes of the form $c_k = c/k^\gamma$ where $c>0$ and $\gamma \in (0,1)$ will be used in all algorithms, and all (function F) measurement noise terms are assumed to have zero means and variances uniformly bounded by $\sigma^2 > 0$. In particular, the constant γ in c_k determines the orders of the estimation bias and variance in a finite-difference scheme, and its choice depends on the step-size parameter α in order to satisfy Assumption A5.

4.1. KW Algorithm

In the KW algorithm, a symmetric finite-difference estimator for $\boldsymbol{G}(\boldsymbol{\theta}_k)$ takes the following form:

$$g_{k,i}(\boldsymbol{\theta}_k) = \frac{F(\boldsymbol{\theta}_k + c_k \boldsymbol{e}_i) - F(\boldsymbol{\theta}_k - c_k \boldsymbol{e}_i)}{2c_k} + \frac{\varepsilon^+_{k,i} - \varepsilon^-_{k,i}}{2c_k}, \quad i = 1, \ldots, d, \tag{5}$$

where $g_{k,i}$ is the ith component of the gradient estimator, $\boldsymbol{e}_i$ denotes the unit vector in the ith direction, and $\varepsilon^\pm_{k,i}$ are the two measurement noise terms when observing F at the perturbed vectors $\boldsymbol{\theta}_k \pm c_k \boldsymbol{e}_i$. Because the noise terms could be correlated, we in general have $E[\varepsilon^+_{k,i} - \varepsilon^-_{k,i} | \mathcal{F}_k] = 0$ and $E[(\varepsilon^+_{k,i} - \varepsilon^-_{k,i})^2 | \mathcal{F}_k] \le 4\sigma^2$ for all $i = 1, \ldots, d$. Note that each iteration of the KW procedure requires $2d$ measurements of the objective function.

Using Taylor's theorem, it is then a simple exercise to show that $E[\|\boldsymbol{b}_k\|^2] \le dc_k^4 L^2/36$ and $E[\|\boldsymbol{\epsilon}_k\|^2] \le d\sigma^2/c_k^2$. Thus, given the form of $c_k = c/k^\gamma$, Assumption A5 is satisfied with $C_1 = dL^2c^4/36$, $q_1 = 2\gamma$ and $C_2 = d\sigma^2/c^2$, $q_2 = \gamma$ provided that $\gamma < \alpha/2$. Consequently, when higher-order terms are ignored, the performance bound in (3) becomes

$$\frac{\sqrt{d}L}{3\rho} c_k^2 + \frac{\sqrt{d}\sigma}{\sqrt{\rho}} \frac{a_k^{1/2}}{c_k}. \tag{6}$$

4.2. Random Direction Finite-Difference Algorithm

A random direction method simultaneously varies all components of the underlying parameter vector in random directions, so that the similar effect of the deterministic finite-difference scheme can be achieved with only two function measurements. The symmetric finite-difference version of the gradient estimator is given by

$$\boldsymbol{g}_k(\boldsymbol{\theta}_k) = \frac{F(\boldsymbol{\theta}_k + c_k \boldsymbol{u}_k) - F(\boldsymbol{\theta}_k - c_k \boldsymbol{u}_k)}{2c_k} \boldsymbol{u}_k + \frac{\varepsilon^+_k - \varepsilon^-_k}{2c_k} \boldsymbol{u}_k, \tag{7}$$

where we take $\boldsymbol{u}_k$ to be the standard normal vector with independent components and assume that $\boldsymbol{u}_k$ is independent of $\varepsilon^\pm_k$.

Through a Taylor series expansion up to the third order, the rth element of the bias term equals

$$b_{k,r} = \frac{c_k^2}{12} E\left[\sum_{i,j,l}^{d} u_{k,i} u_{k,j} u_{k,l} [\nabla^3_{\boldsymbol{\theta}} F(\overline{\boldsymbol{\theta}}^+_k) + \nabla^3_{\boldsymbol{\theta}} F(\overline{\boldsymbol{\theta}}^-_k)]_{i,j,l} u_{k,r} \,\middle|\, \mathcal{F}_k \right], \tag{8}$$

where $\overline{\boldsymbol{\theta}}^\pm_k$ are on the line segments connecting $\boldsymbol{\theta}_k$ and $\boldsymbol{\theta}_k \pm c_k \boldsymbol{u}_k$. Under the current assumptions, a direct application of the Cauchy-Schwarz inequality may lead to a bound for $|b_{k,r}|$ that is too loose, because $[\nabla^3_{\boldsymbol{\theta}} F(\overline{\boldsymbol{\theta}}^\pm_k)]_{i,j,l}$ and $u_{k,i}$'s are not independent. Therefore, we further assume that F is four-times differentiable (with uniformly bounded fourth-order derivatives), so that $b_{k,r}$ can be written as

$$b_{k,r} = \frac{c_k^2}{6} E\left[\sum_{i,j,l}^{d} u_{k,i} u_{k,j} u_{k,l} [\nabla^3_{\boldsymbol{\theta}} F(\boldsymbol{\theta}_k)]_{i,j,l} u_{k,r} \,\middle|\, \mathcal{F}_k \right] + o(c_k^2).$$

Thus, because $\boldsymbol{\theta}_k$ is $\mathcal{F}_k$-measurable, $E[u^2_{k,i}] = 1$, and $E[u^4_{k,i}] = 3$ for all $i = 1, \ldots, d$, a bound on $|b_{k,r}|$ can be derived as follows:

$$\begin{aligned} |b_{k,r}| &\le \frac{c_k^2 L}{6} \sum_{i,j,l}^{d} |E[u_{k,i} u_{k,j} u_{k,l} u_{k,r} | \mathcal{F}_k]| + o(c_k^2) \\ &= \frac{c_k^2 L}{6} \left[3 \sum_{i \ne r} E[u^2_{k,r} u^2_{k,i} | \mathcal{F}_k] + E[u^4_{k,r} | \mathcal{F}_k] \right] + o(c_k^2) \\ &= \frac{dLc_k^2}{2} + o(c_k^2), \end{aligned}$$

implying that $E[\|\boldsymbol{b}_k\|^2] \le d^3 L^2 c_k^4/4 + o(c_k^4)$.

To calculate an upper bound for $E[\|\boldsymbol{\epsilon}_k\|^2]$, we note that

$$\begin{aligned} &E[\|\boldsymbol{g}_k(\boldsymbol{\theta}_k)\|^2 | \mathcal{F}_k] \\ &= E\left[\left(\frac{F(\boldsymbol{\theta}_k + c_k \boldsymbol{u}_k) - F(\boldsymbol{\theta}_k - c_k \boldsymbol{u}_k)}{2c_k} \right)^2 \|\boldsymbol{u}_k\|^2 \,\middle|\, \mathcal{F}_k \right] \\ &\quad + E\left[\frac{(\varepsilon^+_k - \varepsilon^-_k)^2}{4c_k^2} \|u_k\|^2 \,\middle|\, \mathcal{F}_k \right] \\ &\le E\left[\left(\frac{F(\theta_k + c_k u_k) - F(\theta_k - c_k u_k)}{2c_k} \right)^2 \|u_k\|^2 \,\middle|\, \mathcal{F}_k \right] \\ &\quad + \frac{\sigma^2}{c_k^2} E[\|u_k\|^2 | \mathcal{F}_k]. \end{aligned} \tag{9}$$

A second-order Taylor series expansion yields

$$F(\boldsymbol{\theta}_k + c_k \boldsymbol{u}_k) - F(\boldsymbol{\theta}_k - c_k \boldsymbol{u}_k) = 2c_k \boldsymbol{G}^T(\boldsymbol{\theta}_k) \boldsymbol{u}_k + \frac{c_k^2}{2} \boldsymbol{u}_k^T (\boldsymbol{H}(\overline{\boldsymbol{\theta}}^+_k) - \boldsymbol{H}(\overline{\boldsymbol{\theta}}^-_k)) \boldsymbol{u}_k,$$

where $\overline{\boldsymbol{\theta}}^\pm_k$ are on the line segments between $\boldsymbol{\theta}_k$ and $\boldsymbol{\theta}_k \pm c_k \boldsymbol{u}_k$. Substituting the above into (9) and using the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned} E[\|\boldsymbol{g}_k(\boldsymbol{\theta}_k)\|^2 | \mathcal{F}_k] \le E\Big[&(\boldsymbol{G}^T(\boldsymbol{\theta}_k) \boldsymbol{u}_k)^2 \|\boldsymbol{u}_k\|^2 \\ &+ \frac{c_k^2}{16} (\boldsymbol{u}_k^T (\boldsymbol{H}(\overline{\boldsymbol{\theta}}^+_k) - \boldsymbol{H}(\overline{\boldsymbol{\theta}}^-_k)) \boldsymbol{u}_k)^2 \|\boldsymbol{u}_k\|^2 | \mathcal{F}_k \Big] \\ &+ E\Big[\frac{c_k}{2} \|\boldsymbol{G}(\boldsymbol{\theta}_k)\| \|\boldsymbol{u}_k^T (\boldsymbol{H}(\overline{\boldsymbol{\theta}}^+_k) \\ &- \boldsymbol{H}(\overline{\boldsymbol{\theta}}^-_k)) \boldsymbol{u}_k\| \|\boldsymbol{u}_k\|^3 | \mathcal{F}_k \Big] + \frac{\sigma^2}{c_k^2} E[\|\boldsymbol{u}_k\|^2 | \mathcal{F}_k]. \end{aligned}$$

By the Rayleigh-Ritz inequality, $|\boldsymbol{u}_k^T(\boldsymbol{H}(\overline{\boldsymbol{\theta}}_k^+)-\boldsymbol{H}(\overline{\boldsymbol{\theta}}_k^-))\boldsymbol{u}_k| \le (M-m)\|\boldsymbol{u}_k\|^2$. In addition, it is straightforward to verify that $E[(\boldsymbol{G}^T(\boldsymbol{\theta}_k)\boldsymbol{u}_k)^2\|\boldsymbol{u}_k\|^2|\mathcal{F}_k] = (d+2)\|\boldsymbol{G}(\boldsymbol{\theta}_k)\|^2$. Thus, we can further simplify the above bound to get

$$\begin{aligned}
E[\|\boldsymbol{g}_k(\boldsymbol{\theta}_k)\|^2|\mathcal{F}_k] &\le (d+2)\|\boldsymbol{G}(\boldsymbol{\theta}_k)\|^2 \\
&\quad + \frac{c_k}{2}\|\boldsymbol{G}(\boldsymbol{\theta}_k)\|(M-m)E[\|\boldsymbol{u}_k\|^5|\mathcal{F}_k] \\
&\quad + \frac{c_k^2}{16}(M-m)^2E[\|\boldsymbol{u}_k\|^6|\mathcal{F}_k] \\
&\quad + \frac{\sigma^2}{c_k^2}E[\|\boldsymbol{u}_k\|^2|\mathcal{F}_k] \\
&= (d+2)\|\boldsymbol{G}(\boldsymbol{\theta}_k)\|^2 \\
&\quad + \frac{c_k}{2}(M-m)\|\boldsymbol{G}(\boldsymbol{\theta}_k)\|E[Z^{5/2}] \\
&\quad + \frac{c_k^2}{16}(M-m)^2E[Z^3] + \frac{\sigma^2}{c_k^2}E[Z] \\
&\le d(d+2)L_G^2 + \frac{c_k}{2}(M-m)d(d+1)(d+3)L_G \\
&\quad + \frac{c_k^2}{16}(M-m)^2d(d+2)(d+4) + \frac{\sigma^2}{c_k^2}d,
\end{aligned} \tag{10}$$

where we have defined $Z = \|\boldsymbol{u}_k\|^2$, a chi-square random variable with d-degrees of freedom, whose n-th moment is $E[Z^n] = 2^n\Gamma(d/2+n)/\Gamma(d/2)$ with Γ being the gamma function. Note that $E[\|\boldsymbol{\epsilon}_k\|^2|\mathcal{F}_k] = E[\|\boldsymbol{g}_k(\boldsymbol{\theta}_k)\|^2|\mathcal{F}_k] - \|E[\boldsymbol{g}_k(\boldsymbol{\theta}_k)|\mathcal{F}_k]\|^2$. It can be shown that the leading term of $\|E[\boldsymbol{g}_k(\boldsymbol{\theta}_k)|\mathcal{F}_k]\|^2$ is $\|\boldsymbol{G}(\boldsymbol{\theta}_k)\|^2$, which is smaller than dL_G^2. Thus, the right-hand side of (10) serves as a reasonable upper bound for $E[\|\boldsymbol{\epsilon}_k\|^2]$.

Compared with the KW algorithm, it is easy to see from (10) that a random direction method inflates the gradient variance. Nevertheless, the growth rate of the variance is dominated by $\sigma^2 d/c_k^2$, the same as that of KW. Consequently, Assumption A5 is satisfied with $\gamma < \alpha/2$, and according to Theorem 1, when only leading terms are considered, a performance bound on the random direction method is

$$\frac{d^{3/2}L}{\rho}c_k^2 + \frac{\sqrt{d}\sigma}{\sqrt{\rho}}\frac{a_k^{1/2}}{c_k}. \tag{11}$$

4.3. SPSA

The simultaneous perturbation method estimates the gradient $\boldsymbol{G}(\boldsymbol{\theta}_k)$ by

$$\boldsymbol{g}_k(\boldsymbol{\theta}_k) = \frac{F(\boldsymbol{\theta}_k + c_k\boldsymbol{\Delta}_k) - F(\boldsymbol{\theta}_k - c_k\boldsymbol{\Delta}_k)}{2c_k\boldsymbol{\Delta}_k} + \frac{\varepsilon_k^+ - \varepsilon_k^-}{2c_k\boldsymbol{\Delta}_k}, \tag{12}$$

where $\boldsymbol{\Delta}_k = (\Delta_{k,1},\ldots,\Delta_{k,d})^T$ is a zero mean random direction with i.i.d. components.

The most commonly adopted choice of $\boldsymbol{\Delta}_k$ is the symmetric Bernoulli random direction, that is, $P(\Delta_{k,i}=1) = P(\Delta_{k,i}=-1) = 1/2$ for all $i = 1,\ldots,d$. Clearly, because $\Delta_{k,i} \in \{-1,1\}$, we have $1/\boldsymbol{\Delta}_k = \boldsymbol{\Delta}_k$. So, in this setting, (12) becomes a special case of the random direction method (7). Consequently, its bias and variance analysis can be carried out along the same line as in Section 4.2. In particular, by noting that $E[\Delta_{k,i}\Delta_{k,j}] = 0$ for $i \ne j$, and $\Delta_{k,i}^2 = 1$ for all $i = 1,\ldots,d$, it is easy to observe that $E[\|\boldsymbol{b}_k\|^2] \le d^3L^2c_k^4/4 + o(c_k^4)$.

On the other hand, by noting that $\|\boldsymbol{\Delta}_k\|^2 = d$, we obtain from (9) that

$$E[\|\boldsymbol{g}_k(\boldsymbol{\theta}_k)\|^2|\mathcal{F}_k] \le dE\left[\left(\frac{F(\boldsymbol{\theta}_k + c_k\boldsymbol{\Delta}_k) - F(\boldsymbol{\theta}_k - c_k\boldsymbol{\Delta}_k)}{2c_k}\right)^2\middle|\mathcal{F}_k\right] + \frac{d\sigma^2}{c_k^2}. \tag{13}$$

Again, by a Taylor series expansion,

$$\begin{aligned}
F(\boldsymbol{\theta}_k + c_k\boldsymbol{\Delta}_k) - F(\boldsymbol{\theta}_k - c_k\boldsymbol{\Delta}_k) &= 2c_k\boldsymbol{G}^T(\boldsymbol{\theta}_k)\boldsymbol{\Delta}_k \\
&\quad + \frac{c_k^2}{2}\boldsymbol{\Delta}_k^T(\boldsymbol{H}(\overline{\boldsymbol{\theta}}_k^+) - \boldsymbol{H}(\overline{\boldsymbol{\theta}}_k^-))\boldsymbol{\Delta}_k,
\end{aligned}$$

where $\overline{\boldsymbol{\theta}}_k^{\pm}$ are on the line segments between $\boldsymbol{\theta}_k$ and $\boldsymbol{\theta}_k \pm c_k\boldsymbol{\Delta}_k$. Substituting the above into (13) and using the inequality $|\boldsymbol{\Delta}_k^T(\boldsymbol{H}(\overline{\boldsymbol{\theta}}_k^-) - \boldsymbol{H}(\overline{\boldsymbol{\theta}}_k^-))\boldsymbol{\Delta}_k| \le (M-m)\|\boldsymbol{\Delta}_k\|^2 = d(M-m)$ and the fact $E[(\boldsymbol{G}^T(\boldsymbol{\theta}_k)\boldsymbol{\Delta}_k)^2|\mathcal{F}_k] = \|\boldsymbol{G}(\boldsymbol{\theta}_k)\|^2$, we finally obtain

$$\begin{aligned}
E[\|\boldsymbol{g}_k(\boldsymbol{\theta}_k)\|^2|\mathcal{F}_k] &\le d\|\boldsymbol{G}(\boldsymbol{\theta}_k)\|^2 + \frac{d^2\sqrt{d}c_k}{2}\|\boldsymbol{G}(\boldsymbol{\theta}_k)\|(M-m) \\
&\quad + \frac{c_k^2}{16}d^3(M-m)^2 + \frac{d\sigma^2}{c_k^2} \\
&\le d^2L_G^2 + \frac{d^3L_G(M-m)}{2}c_k \\
&\quad + \frac{d^3(M-m)^2}{16}c_k^2 + \frac{d\sigma^2}{c_k^2}.
\end{aligned} \tag{14}$$

Finally, by (3), a bound on the performance of SPSA, when expressed in terms of its leading terms, is given by

$$\frac{d^{3/2}L}{\rho}c_k^2 + \frac{\sqrt{d}\sigma}{\sqrt{\rho}}\frac{a_k^{1/2}}{c_k}. \tag{15}$$

4.4. Observations

In view of the performance bounds given by (6), (11), and (15), we have the following observations:

(i) Consider the case $\alpha = 6\gamma$, that is, both c_k^2 and $a_k^{1/2}/c_k$ have the same order $O(k^{-\alpha/3})$, which yields the best convergence rate for a given a_k. Suppose that both KW and random direction are implemented using the same a_k and c_k, and that the choice of constant c in c_k does not depend on or vary with σ, which is assumed unknown here. Then a comparison of (6) and (11) indicates that if the noise level σ is large compared with the problem dimension d (i.e., the $\sqrt{d}\sigma/\sqrt{\rho}$ terms in (6) and (11) dominate), then using a random direction method

will likely result in a significant improvement in algorithm efficiency, especially for higher-dimensional problems. On the other hand, when there is no noise or the noise variance is very small, the $d^{3/2}L/\rho$ (bias) term in the bound (11) dominates, and it is easy to observe that a random direction-based algorithm would roughly take $d^{3/\alpha} \geq d^3$ (because $\alpha < 1$) times the number of iterations required by KW in order to achieve the same level of accuracy, and this clearly nullifies the benefit of the d-fold reduction in the number of function evaluations at each iteration. However, there might be some exceptions, for example, when the third-order derivatives of F are very small (e.g., $L = 0$ for quadratic functions) so that the bias terms vanish, or when the mixed third-order partial derivatives of F are all zeros, in which case it can be seen from (8) that the dependency of the bias terms in (11) and (15) on the problem dimension reduces from $d^{3/2}$ to $\sqrt{d}$.

The above comparison assumes that no problem-specific information is available. When the noise level σ either is known or can be reliably estimated, that is, the setting examined in Berahas et al. (2022), the constant c in the perturbation-size c_k can further be optimized based on σ, leading to $c = (3\sqrt{\rho}\sigma/(2L))^{1/3}a^{1/6}$ for KW and $c = (\sqrt{\rho}\sigma/(2dL))^{1/3}a^{1/6}$ for random direction. It is not difficult to show that the bounds in (6) and (11) then scale respectively as $O(d^{1/2})\sigma^{2/3}k^{-\alpha/3}$ and $O(d^{5/6})\sigma^{2/3}k^{-\alpha/3}$; both have the same order of dependency on σ. Because $d^{1/2}k^{-\alpha/3} = d^{5/6}(d^{1/\alpha}k)^{-\alpha/3}$, this means that regardless of the noise level σ, a random direction method would require at least $d^{1/\alpha} > d$ (because $\alpha < 1$) times number of iterations in order to attain the same level of accuracy as KW. *In other words, if the perturbation-size in KW is allowed to be chosen optimally using knowledge of σ, then the algorithm cannot be outperformed by a random direction method.* This observation is consistent with the findings of Scheinberg (2022) and Berahas et al. (2022) in the constant step-size setting.

(ii) When $2\gamma < \alpha < 6\gamma$, that is, $c_k^2 = o(a_k^{1/2}/c_k)$, the influence of the problem dimension on the bias will eventually damp out as k increases. Thus, assuming that the same α_k and c_k are used, a random direction method could be advantageous, especially when the noise variance is high.

(iii) When $\alpha > 6\gamma$, then the bias term dominates, so KW may in general yield superior performance in the long run, at least in theory. In particular, when there is no estimation noise, our performance bounds indicate that the number of iterations required by a random direction-based algorithm is more than d^3 times that of KW.

(iv) When only leading terms are considered, (11) and (15) are identical. Thus, the performance of random direction and SPSA should be similar. However, because of the use of different random directions, a comparison of (10) and (14) indicates that the dependencies on problem dimension in the nonleading terms are $d(d+2)$, $d(d+1)(d+2)$, and $d(d+2)(d+4)$ in random direction versus d^2 and d^3 in SPSA. This may have a nonnegligible influence on the algorithm performance when k is small.

(v) When constant step- and perturbation-sizes are used, both c_k^2 and $a_k^{1/2}/c_k$ are of order $O(1)$. This can be viewed as a special case of Case (i) with $\alpha = \gamma = 0$, except that the bounds on gradient variances in (10) and (14) may no longer be dominated by $d\sigma^2/c^2$. However, irrespective of the actual dominating terms in (10) and (14), the performance bounds for random direction and SPSA are always worse (larger) than those given by (11) and (15). Therefore, assuming that the conditions of Corollary 1 are met, essentially the same conclusions as in Case (i) above can be made.

5. Conclusions

For a class of problems with convex differentiable structures, we have established a finite-time performance bound for gradient descent algorithms under general conditions on the gradient estimation errors. The bound allows for a detailed characterization of an algorithm's rate of convergence through directly analyzing the bias and variance of the gradient estimator when employed in an iterative search. Two types of finite-difference-based gradient approximation methods, deterministic FD and random direction FD, are then studied and compared in terms of their efficiency based on the derived bound.

An open question is whether the typical d-fold per-iteration reduction in the number of performance evaluations of a random direction method will justify the potential increase in the number of gradient descent iterations. Prior studies on this topic are primarily based on asymptotic theory (Spall 1992, Kushner and Yin 1997) or by means of one-step bias-variance analysis (Berahas et al. 2022, Scheinberg 2022). Our finite-time study allows us to compare algorithm performance under different parameter settings, providing a more thorough understanding of this issue. A case of interest is when a_k and c_k take the forms $a_k = a/k^\alpha$ and $c_k = c/k^\gamma$ with $\alpha = 6\gamma$. In particular, if the selection of c does not depend on the noise level σ, then our analysis indicates that the relative efficiency of a random direction method is generally contingent upon the amount of variability in the measurement noise in relation to the problem dimension. On the other hand, in the setting where the knowledge of σ can be exploited to optimize the choice of c, we obtain the negative result that the performance of a deterministic FD-based algorithm in general cannot be further improved through the use of random direction methods, an observation that is in agreement with earlier results reported in Scheinberg (2022) and Berahas et al. (2022). The essence of the issue is that when compared with deterministic FD estimators, existing randomized gradient estimators would lead to an extra d-fold increase in

the estimation bias, which plays a major role in the bias-variance trade-off.

Acknowledgments

The authors thank the editors and two anonymous referees for their helpful comments and suggestions that have led to a substantially improved note.

Appendix. Proof of Lemma 1

Let $u(x) = s/x^p$, $r(x) = 1/x^q$, and $U(x) = \int_1^x u(y)dy$ for $x > 0$. We can find a constant $C > 0$ and an integer $N > 0$ such that $u(i) < 1$, $w(i) \le Ci^{-q} = Cr(i)$ for all $i \ge N$, and that $e^{U(x)}u(x)r(x)$ and $-e^{U(x)}r'(x)$ are both increasing when $x \ge N$. Thus, we have for all $k \ge N$ that

$$\begin{aligned}
&\sum_{i=1}^{k}\left[\prod_{j=i+1}^{k}(1-u(j))\right]u(i)w(i)\\
&=\sum_{i=1}^{N-1}\left[\prod_{j=i+1}^{k}(1-u(j))\right]u(i)w(i)+\sum_{i=N}^{k}\left[\prod_{j=i+1}^{k}(1-u(j))\right]u(i)w(i)\\
&=\prod_{j=N}^{k}(1-u(j))\sum_{i=1}^{N-1}\left[\prod_{j=i+1}^{N-1}(1-u(j))\right]u(i)w(i)+\sum_{i=N}^{k}\left[\prod_{j=i+1}^{k}(1-u(j))\right]u(i)w(i)\\
&\le e^{-\sum_{j=N}^{k}u(j)}\left|\sum_{i=1}^{N-1}\prod_{j=i+1}^{N-1}(1-u(j))u(i)w(i)\right|+\sum_{i=N}^{k}\left[\prod_{j=i+1}^{k}(1-u(j))\right]u(i)w(i)\\
&=O(e^{-\frac{s}{1-p}k^{1-p}})+C\sum_{i=N}^{k}\left[\prod_{j=i+1}^{k}(1-u(j))\right]u(i)r(i). \qquad \text{(A.1)}
\end{aligned}$$

The second term on the right-hand side of (A.1) can be bounded as follows:

$$\begin{aligned}
&\sum_{i=N}^{k}\left[\prod_{j=i+1}^{k}(1-u(j))\right]u(i)r(i)\\
&\le\sum_{i=N}^{k}e^{-\sum_{j=i+1}^{k}u(j)}u(i)r(i)\\
&\le\sum_{i=N}^{k}e^{-\int_{i+1}^{k+1}u(x)dx}u(i)r(i)\\
&=e^{-U(k+1)}\sum_{i=N}^{k}e^{U(i+1)}u(i)r(i)\\
&\le e^{\frac{s}{1-p}}e^{-U(k+1)}\sum_{i=N}^{k}e^{U(i)}u(i)r(i)\\
&\le e^{\frac{s}{1-p}}e^{-U(k+1)}\int_{N}^{k+1}e^{U(x)}u(x)r(x)dx\\
&\le e^{\frac{s}{1-p}}e^{-U(k+1)}\left(r(k+1)e^{U(k+1)}-\int_{N}^{k+1}e^{U(x)}r'(x)dx\right)\\
&\le e^{\frac{s}{1-p}}(r(k+1)-(k+1-N)r'(k+1))\\
&=O(k^{-q}).
\end{aligned}$$

Finally, because $p \in (0,1)$, the proof is hence completed by noting that the first term on the right-hand size of (A.1) is of order $o(k^{-q})$. □

References

Berahas AS, Cao L, Choromanski K, Scheinberg K (2022) A theoretical and empirical comparison of gradient approximations in derivative-free optimization. *Found. Comput. Math.* 22(2):507–560.

Bottou L, Curtis FE, Nocedal J (2018) Optimization methods for large-scale machine learning. *SIAM Rev.* 60(2):223–311.

Chen J, Luss R (2019) Stochastic gradient descent with biased but consistent gradient estimators. Preprint, submitted July 31, 2018, https://arxiv.org/abs/1807.11880.

Demidovich Y, Malinovsky G, Sokolov I, Richtárik R (2023) A guide through the zoo of biased SGD. Preprint, submitted May 25, https://arxiv.org/abs/2305.16296.

Driggs D, Liang J, Schönlieb C (2022) On biased stochastic gradient estimation. *J. Machine Learn. Res.* 23(24):1057–1099.

Duchi JC, Jordan MI, Wainwright MJ, Wibisono A (2015) Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Trans. Inform. Theory* 61(5): 2788–2806.

Ghadimi S, Lan G (2012) Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization I: A generic algorithmic framework. *SIAM J. Optim.* 22(4):1469–1492.

Hu B, Seiler P, Lessard L (2021) Analysis of biased stochastic gradient descent using sequential semidefinite programs. *Math. Programming* 187(1–2):383–408.

Hu J, Song M, Fu M (2024) Quantile optimization via multiple time-scale local search for black-box functions. *Oper. Res.* Forthcoming.

Karimi B, Miasojedow B, Moulines E, Wai HT (2019) Non-asymptotic analysis of biased stochastic approximation scheme. *Proc. Machine Learn. Res.* 99:1–31.

Kiefer J, Wolfowitz J (1952) Stochastic estimation of the maximum of a regression function. *Ann. Math. Statist.* 23(3):462–466.

Kushner HJ, Yin GG (1997) *Stochastic Approximation and Recursive Algorithms and Applications* (Springer, New York).

Rugh WJ (1996) *Linear System Theory*, 2nd ed. (Prentice Hall, Upper Saddle River, NJ).

Scheinberg K (2022) Finite difference gradient approximation: To randomize or not? *INFORMS J. Comput.* 34(5):2384–2388.

Spall JC (1992) Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Trans. Automat. Control* 37(3):332–341.

Jiaqiao Hu is an associate professor in the Department of Applied Mathematics and Statistics at the State University of New York, Stony Brook. His research interests include Markov decision processes, simulation-based optimization, stochastic modeling and analysis, and computational learning theory.

Michael C. Fu holds the Smith Chair of Management Science in the Robert H. Smith School of Business, with a joint appointment in the Institute for Systems Research and affiliate faculty appointment in the Department of Electrical and Computer Engineering, all at the University of Maryland, College Park. His research interests include Markov decision processes, stochastic gradient estimation, simulation optimization, and applied probability. He is a Fellow of INFORMS and IEEE.