Proceedings of the 2024 Winter Simulation Conference
H. Lam, E. Azar, D. Batur, S. Gao, W. Xie, S. R. Hunter, and M. D. Rossetti, eds.

CENTRAL FINITE-DIFFERENCE BASED GRADIENT ESTIMATION METHODS FOR STOCHASTIC OPTIMIZATION

Raghu Bollapragada[1], Cem Karamanli[1], and Stefan M. Wild[2]

[1]Operations Research and Industrial Engineering, The University of Texas at Austin, Austin, TX, USA
[2]Applied Math. and Comp. Res. Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

ABSTRACT

This paper presents an algorithmic framework for solving unconstrained stochastic optimization problems using only stochastic function evaluations. We employ central finite-difference based gradient estimation methods to approximate the gradients and dynamically control the accuracy of these approximations by adjusting the sample sizes used in stochastic realizations. We analyze the theoretical properties of the proposed framework on nonconvex functions. Our analysis yields sublinear convergence results to the neighborhood of the solution, and establishes the optimal worst-case iteration complexity ($\mathscr{O}(\varepsilon^{-1})$) and sample complexity ($\mathscr{O}(\varepsilon^{-2})$) for each gradient estimation method to achieve an ε-accurate solution. Finally, we demonstrate the performance of the proposed framework and the quality of the gradient estimation methods through numerical experiments on nonlinear least squares problems.

1 INTRODUCTION

We consider unconstrained stochastic optimization problems of the form

$$\min_{x\in\mathbb{R}^d} F(x) = \mathbb{E}_\zeta\left[f(x,\zeta)\right], \tag{1}$$

where $F:\mathbb{R}^d\to\mathbb{R}$ is a continuously differentiable function, ζ is a random variable with associated probability space $(\Xi,\mathscr{F},P)$, $f:\mathbb{R}^d\times\Xi\to\mathbb{R}$, and $\mathbb{E}_\zeta[\cdot]$ denotes the expectation with respect to P. We consider derivative-free optimization (DFO) settings where gradient information is unavailable, and only stochastic realizations of the objective functions, obtained through a zeroth-order oracle, are available. Such problem settings arise in different applications, including simulation optimization (Blanchet et al. 2019; Pasupathy et al. 2018; Shashaani et al. 2018) and reinforcement learning (Bertsekas 2019).

Gradient estimation methods, which involve estimating gradients through function evaluations and incorporating these estimators into optimization algorithms, are widely recognized within DFO settings. Recently, Bollapragada et al. (2024) proposed a unified algorithmic framework that encompasses stochastic gradient estimation methods for solving (1). This framework incorporates forward finite-difference methods for gradient estimation and controls the accuracy of these estimators by utilizing common random number (CRN) settings and adaptively choosing the number of stochastic realizations (sample sizes) employed in these estimators.

In this paper, we extend this framework to include central finite-difference methods, which are an alternative to forward finite-difference methods and have the potential to achieve superior performance under specific settings, such as the standard finite-difference method employed within CRN settings (L'Ecuyer and Yin 1998; Larson et al. 2019). Specifically, we propose iterative algorithms that incorporate gradient estimators computed using subsampled functions defined as

$$F_{S_k}(x) := \frac{1}{|S_k|}\sum_{\zeta_i\in S_k} f(x,\zeta_i) \qquad \forall x\in\mathbb{R}^d, \tag{2}$$

979-8-3315-3420-2/24/$31.00 ©2024

where $S_k = \{\zeta_1, \dots, \zeta_{|S_k|}\}$ is a set of random realizations at each iteration k. By utilizing a set of vectors T_k, we derive the general central finite-difference based gradient estimators within CRN settings as follows

$$g_{S_k,T_k}(x_k) := \gamma_k \sum_{u_j \in T_k} \left(\frac{F_{S_k}(x_k + \nu u_j) - F_{S_k}(x_k - \nu u_j)}{2\nu} \right) u_j, \tag{3}$$

where $\nu > 0$ is the sampling radius, and $\gamma_k > 0$ is the scaling coefficient.

Within this framework, we consider central finite-difference variants of various methods, encompassing standard finite-difference methods (cFD) (Blum 1954; Kiefer and Wolfowitz 1952), Gaussian smoothing methods (cGS) (Nesterov and Spokoiny 2017), sphere smoothing methods (cSS) (Berahas et al. 2021), randomized coordinate finite-difference methods (cRC) (Wright 2015), and randomized subspace finite-difference methods (cRS) (Berahas et al. 2021; Kozak et al. 2021). These methods differ in their choices of T_k and γ_k (Bollapragada et al. 2024, Table 2.1). The u_j vectors are canonical vectors e_j for cFD and cRC, sampled from a multivariate standard normal distribution ($\mathcal{N}(0,I)$) for cGS, drawn from a uniform distribution over the surface of a unit sphere ($\mathcal{U}(\mathcal{S}(0,1))$) for cSS, and random orthonormal vectors for cRS. The value of γ_k is 1 for cFD, $1/|T_k|$ for cGS, and $d/|T_k|$ for cSS, cRC, and cRS.

The general iterative update rule within this framework, employing the gradient estimators defined in (3), is given as

$$x_{k+1} = x_k - \alpha_k g_{S_k,T_k}(x_k), \tag{4}$$

where α_k is the step size. The efficiency of this update rule depends on the accuracy of the gradient estimators. Bollapragada et al. (2024) incorporated an adaptive sampling approach to control the gradient estimation accuracy as the algorithm progresses by adjusting the sample sizes ($|S_k|$) at each iteration. The key idea behind adaptive sampling approaches is that inaccurate gradient approximations are sufficient when the current iterate is far away from the solution, and the accuracy in these approximations should increase as the iterates approach the optimal solution. Such approaches retain the optimal theoretical convergence properties of their deterministic counterparts while being efficient. In this work, we adapt these strategies to the central finite-difference based methods and extend their analysis to nonconvex problem settings.

The paper is organized as follows. A literature review and a summary of our notation are presented in the remainder of Section 1. Mathematical preliminaries, including assumptions and sampling conditions, are discussed in Section 2. In Section 3, we provide theoretical convergence and complexity results, while Section 4 presents the numerical results. Finally, in Section 5, we make some concluding remarks.

1.1 Literature Review

Several gradient estimation methods (Kiefer and Wolfowitz 1952; Nesterov and Spokoiny 2017; Flaxman et al. 2005; Berahas et al. 2021) have been proposed for both deterministic and stochastic optimization, as extensively reviewed in (Conn et al. 2009; Larson et al. 2019). Pasupathy et al. (2018) analyzed the convergence properties of the fundamental gradient estimation method introduced by Kiefer and Wolfowitz (1952) under different sampling rates. Ghadimi and Lan (2013) provided optimal worst-case sample complexities for both convex and nonconvex functions, showing $\mathcal{O}(\varepsilon^{-2})$ complexity to achieve an ε-accurate solution for stochastic approximation based gradient estimation methods, with different accuracy definitions for convex and nonconvex problems. Gradient estimation methods have also been integrated into quasi-Newton approaches (Berahas et al. 2019; Bollapragada and Wild 2023; Marrinan et al. 2023). Additionally, adaptive sampling approaches, well-established in stochastic optimization (Byrd et al. 2012; Bollapragada et al. 2018; Bollapragada et al. 2023; Berahas et al. 2022), have recently been applied in DFO settings (Shashaani et al. 2018; Bollapragada and Wild 2023; Bollapragada et al. 2024). Bollapragada and Wild (2023) generalized the norm condition (Byrd et al. 2012) and the practical inner-product condition (Bollapragada et al. 2018) to standard finite-difference based gradient estimation methods, while Bollapragada et al. (2024) extended these conditions to other forward finite-difference based gradient estimation methods.

1.2 Notation

The set of nonnegative integers and positive integers is denoted by $\mathbb{Z}_+ := \{0,1,2,\dots\}$ and $\mathbb{Z}_{++} := \{1,2,\dots\}$ respectively. The set of real numbers (scalars) is denoted by $\mathbb{R}$, the set of d-dimensional vectors is denoted by $\mathbb{R}^d$, and the set of m-by-d matrices is denoted by $\mathbb{R}^{m\times d}$. The transpose of a matrix $A \in \mathbb{R}^{m\times d}$ is denoted by $A^T \in \mathbb{R}^{d\times m}$. Throughout the paper, the ℓ_2 vector norm or matrix norm is denoted by $\|\cdot\|$.

2 PRELIMINARIES

In this section, we outline the mathematical preliminaries, including the assumptions made throughout the paper and the conditions used to determine the sample sizes in the stochastic approximations at each iteration. We begin by stating the assumptions about the objective function.

Assumption A The objective function $F : \mathbb{R}^d \to \mathbb{R}$ is twice continuously differentiable function and is bounded below. That is, there exists $F^* > -\infty$ such that $F^* := \inf_{x\in\mathbb{R}^d} F(x)$. Furthermore, the stochastic function $f(\cdot,\zeta) : \mathbb{R}^d \to \mathbb{R}$ is also twice continuously differentiable with Lipschitz continuous gradients and Hessians with Lipschitz constants $L_f < \infty$ and $M_f < \infty$, respectively. That is, for every ζ,

$$\|\nabla f(x,\zeta) - \nabla f(y,\zeta)\| \le L_f\|x-y\|, \quad \text{and} \quad \|\nabla^2 f(x,\zeta) - \nabla^2 f(y,\zeta)\| \le M_f\|x-y\| \quad \forall x,y \in \mathbb{R}^d.$$

Assumption A implies that the objective function F has Lipschitz continuous gradients and Hessians with Lipschitz constants $L_F \le L_f$ and $M_F \le M_f$, respectively, which is a common assumption in central finite-difference settings (Berahas et al. 2021). While it is possible to relax the assumption on the smoothness of stochastic functions and require only F to be smooth (Bollapragada and Wild 2023), such assumptions on the stochastic functions are useful in providing the complexity analysis (Bollapragada et al. 2024). Moreover, under Assumption A, it follows from the second-order Taylor expansion that for every $\nu > 0$, ζ, and $x,u \in \mathbb{R}^d$,

$$f(x+\nu u,\zeta) - f(x-\nu u,\zeta) \le 2\nu u^T\nabla f(x,\zeta) + \frac{M_f}{3}\nu^3\|u\|^3, \tag{5}$$

$$F(x+\nu u) - F(x-\nu u) \le 2\nu u^T\nabla F(x) + \frac{M_F}{3}\nu^3\|u\|^3. \tag{6}$$

We also assume that the variance in the stochastic gradients is bounded, a standard assumption in stochastic optimization literature (Bottou et al. 2018).

Assumption B There exist constants $\beta_1,\beta_2 \ge 0$ such that

$$\mathbb{E}_{\zeta_i}[\|\nabla f(x,\zeta_i) - \nabla F(x)\|^2] \le \beta_1\|\nabla F(x)\|^2 + \beta_2, \quad \forall x \in \mathbb{R}^d.$$

The next assumption pertains to the independence of sets S_k and T_k sampled at each iteration k.

Assumption C At every iteration k, the sample set S_k consists of independent and identically distributed (i.i.d.) samples of ζ. That is, for all $x \in \mathbb{R}^d$ and $k \in \mathbb{Z}_+$,

$$\mathbb{E}_{\zeta_i}[f(x,\zeta_i)] = F(x), \qquad \forall \zeta_i \in S_k.$$

Moreover, the vector set T_k comprises sampled vectors that are chosen independently of the sample set S_k.

We also define the conditional expectations with respect to these random subsets, which are the only sources of randomness in the iterates generated by (4). That is, we define the filtrations $\mathscr{F}_k = \sigma(x_0,\{T_1,S_1\},\{T_2,S_2\},\cdots,\{T_{k-1},S_{k-1}\})$ and $\mathscr{F}_{k+1/2} = \sigma(x_0,\{T_1,S_1\},\{T_2,S_2\},\cdots,\{T_{k-1},S_{k-1}\},T_k)$,

$$\mathbb{E}_k[\cdot] = \mathbb{E}_{T_k}[\cdot] := \mathbb{E}[\cdot|\mathscr{F}_k], \quad \text{and} \quad \mathbb{E}_{S_k}[\cdot] := \mathbb{E}\left[\cdot\middle|\mathscr{F}_{k+1/2}\right].$$

Moreover, we define the following expected quantities

$$g_{T_k}(x_k) := \mathbb{E}_{S_k}[g_{S_k,T_k}(x_k)], \quad \text{and} \quad g(x_k) := \mathbb{E}_{T_k}[g_{T_k}(x_k)]. \tag{7}$$

Using these definitions, we can decompose the error in the gradient estimator into three terms:

$$g_{S_k,T_k}(x_k) - \nabla F(x_k) = \underbrace{g_{S_k,T_k}(x_k) - g_{T_k}(x_k)}_{\text{function sampling error}} + \underbrace{g_{T_k}(x_k) - g(x_k)}_{\text{vector sampling error}} + \underbrace{g(x_k) - \nabla F(x_k),}_{\text{bias}}$$

where the function sampling error depends on the choice of S_k, vector sampling error depends on the choice of T_k, and the bias term, arising from the absence of gradient information, depends on the choice of ν. While we choose a constant $|T_k|$ and ν throughout the iterations, $|S_k|$ is adaptively chosen to control the function sampling error, thereby controlling the gradient estimation error. Guided by this principle, Bollapragada et al. (2024) proposed the following theoretical condition that generalizes the well-known *norm condition* in the derivative-based methods (Byrd et al. 2012; Bollapragada et al. 2018).

Condition 1 (Theoretical Norm Condition) (Bollapragada et al. 2024, Condition 3)

$$\frac{\mathbb{E}_{T_k}[\mathbb{E}_{\zeta_i}[\|g_{\zeta_i,T_k}(x_k) - g_{T_k}(x_k)\|^2]]}{|S_k|} \leq \theta^2 \mathbb{E}_{T_k}[\|g_{T_k}(x_k)\|^2], \quad \theta > 0. \tag{8}$$

We choose the sample sizes $|S_k|$ at each iteration such that this condition is satisfied. Evaluating this condition requires computing population (expectation) quantities that may not be available in practice, and we provide a practical test that employs sampled quantities to overcome this limitation in Section 4.

3 THEORETICAL RESULTS

In this section, we provide theoretical convergence guarantees and worst-case complexity results for the iterates generated by (4) with sample sizes $|S_k|$ satisfying Condition 1.

3.1 Convergence Results

We first establish a descent lemma that provides an upper bound on the expected decrease in the function value per iteration.

Lemma 1 For any $x_0 \in \mathbb{R}^d$, let $\{x_k : k \in \mathbb{Z}_{++}\}$ be generated by iteration (4) with $|S_k|$ satisfying Condition 1 for a given constant $\theta > 0$. Suppose that Assumptions A, B, and C hold. Then, for any $k \in \mathbb{Z}_+$, if α_k satisfies

$$0 < \alpha_k \leq \bar{\alpha}_k, \tag{9}$$

we have

$$\mathbb{E}_k[F(x_{k+1})] \leq F(x_k) - \frac{\alpha_k}{4}\|\nabla F(x_k)\|^2 + \alpha_k \chi_k \quad \forall k \in \mathbb{Z}_+, \tag{10}$$

where $\bar{\alpha}_k, \chi_k > 0$ are given in Table 1.

Proof. Using the result of Bollapragada et al. (2024), Lemma 3.2, it follows that

$$\mathbb{E}_k[F(x_{k+1})] \leq F(x_k) + \frac{\alpha_k}{2}\|\delta_k\|^2 - \frac{\alpha_k}{2}\|\nabla F(x_k)\|^2$$
$$+ L_F \alpha_k^2 (1+\theta^2)\left(\|\delta_k\|^2 + \|\nabla F(x_k)\|^2 + \frac{1}{2}(\mathbb{E}_{T_k}[\|g_{T_k}(x_k) - g(x_k)\|^2])\right),$$

where $\delta_k := g(x_k) - \nabla F(x_k)$. The rest of the proof follows by utilizing upper bounds $\bar{\delta}$ on δ_k given in Table 1 and analyzing the variance terms $(\mathbb{E}_{T_k}[\|g_{T_k}(x_k) - g(x_k)\|^2])$ for each method individually. The $\bar{\delta}$ values corresponding to cFD, cGS, and cSS methods, as well as the upper bounds for the variances of cGS and cSS, are provided in Berahas et al. (2021). For the remaining methods, these values could be obtained by following a similar procedure outlined in Bollapragada et al. (2024), Lemma 3.3. □

Lemma 1 demonstrates that the function values are expected to decrease at any iteration k, provided the gradient ($\|\nabla F(x_k)\|^2$) is sufficiently large. However, the term χ_k hinders the progress. In fact, when a fixed step size ($\alpha_k = \bar{\alpha}$) and a fixed number of directions ($|T_k| = N$) are employed, the χ_k term becomes constant, and the iterates converge to a neighborhood of the solution. The following theorem establishes the convergence rate and the size of the neighborhood for nonconvex functions.

Theorem 1 (Sublinear Convergence). Suppose the conditions for Lemma 1 hold. Let $\bar{\alpha}$ and $\bar{\chi}$ be the values obtained by choosing $|T_k| = N$ in Table 1. If $\alpha_k = \bar{\alpha}$, for all $k \in \mathbb{Z}_+$, then the sequence $\{x_k : k \in \mathbb{Z}_+\}$ converges sublinearly to a neighborhood of the solution. That is,

$$\min_{0\le k\le K-1} \mathbb{E}[\|\nabla F(x_k)\|^2] \le \frac{4(F(x_0) - F^*)}{K\bar{\alpha}} + 4\bar{\chi}. \tag{11}$$

Moreover, for any $p \in (0,1]$, we have with probability at least $1-p$ that

$$\min_{0\le k\le K-1} \|\nabla F(x_k)\|^2 \le \frac{4(F(x_0) - F^*)}{K\bar{\alpha}p} + \frac{4\bar{\chi}}{p}. \tag{12}$$

Proof. Rearranging (10) yields

$$\|\nabla F(x_k)\|^2 \le \frac{4(F(x_k) - \mathbb{E}_k[F(x_{k+1})])}{\bar{\alpha}} + 4\bar{\chi}.$$

Taking the total expectation and averaging the terms from $k = 0$ to $k = K-1$ yields

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}[\|\nabla F(x_k)\|^2] \le \frac{1}{K}\sum_{k=0}^{K-1}\left[\frac{4\mathbb{E}[F(x_k) - F(x_{k+1})]}{\bar{\alpha}} + 4\bar{\chi}\right] = \frac{4\mathbb{E}[F(x_0) - F(x_K)]}{K\bar{\alpha}} + 4\bar{\chi}. \tag{13}$$

Using the above inequality, along with the fact that $\min_{0\le k\le K-1} \mathbb{E}[\|\nabla F(x_k)\|^2] \le \frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}[\|\nabla F(x_k)\|^2]$, and $F(x_K) \ge F^*$ due to Assumption A, we obtain (11). Moreover, using the fact that $\min_{0\le k\le K-1} \|\nabla F(x_k)\|^2 \le \frac{1}{K}\sum_{k=0}^{K-1} \|\nabla F(x_k)\|^2$, we get

$$\begin{aligned}
P\left(\min_{0\le k\le K-1} \|\nabla F(x_k)\|^2 > \frac{4(F(x_0) - F^*)}{K\bar{\alpha}p} + \frac{4\bar{\chi}}{p}\right) &\le P\left(\frac{1}{K}\sum_{k=0}^{K-1} \|\nabla F(x_k)\|^2 > \frac{4(F(x_0) - F^*)}{K\bar{\alpha}p} + \frac{4\bar{\chi}}{p}\right) \\
&\le \frac{\mathbb{E}[\frac{1}{K}\sum_{k=0}^{K-1} \|\nabla F(x_k)\|^2]}{\frac{4(F(x_0)-F^*)}{K\bar{\alpha}p} + \frac{4\bar{\chi}}{p}} \\
&\le \frac{\frac{4\mathbb{E}[F(x_0)-F(x_K)]}{K\bar{\alpha}} + 4\bar{\chi}}{\frac{4(F(x_0)-F^*)}{K\bar{\alpha}p} + \frac{4\bar{\chi}}{p}} \le p.
\end{aligned}$$

where the second inequality is due to Markov's inequality, the third inequality is due to (13), and the last inequality is due to the fact that $F(x_K) \ge F^*$, which completes the proof. □

Theorem 1 provides convergence results both in expectation and in probability. The difference between different gradient estimation methods in terms of convergence behavior is reflected in the step size $\bar{\alpha}$ and neighborhood $\bar{\chi}$ parameters (see Table 1). We note that similar observations to those made in forward finite-difference methods (Bollapragada et al. 2024) are found here. cRC and cRS methods have convergence rates $\frac{d}{N}$ times worse compared to cFD. Similarly, convergence rates of cGS and cSS methods are $\frac{N+4.5d}{N}$ times worse than that of cFD. Regarding the size of convergence neighborhoods, we observe that cFD, cRC, and cRS methods have the same neighborhood size. Furthermore, assuming that N is small and d is large, cSS has a neighborhood size similar to cFD, whereas cGS has a neighborhood size that is d^2 times larger compared to cFD.

Table 1: Properties and convergence results for different gradient estimation methods. $\bar{\delta}$ serves as an upper bound on $\|\delta_k\|$, where $\delta_k := g(x_k) - \nabla F(x_k)$. $\bar{\alpha}_k$ and χ_k values summarize the results of Lemma 1. The $\nu = \hat{\nu}$ values ensure that the convergence neighborhood $4\bar{\chi}$ (as stated in Theorem 1) equals $\frac{\varepsilon p}{2}$. Here, $\Omega_1 := \frac{1}{4(1+\theta^2)L_F}$, $\Omega_2 := \frac{dM_F^2\nu^4}{48}$, and $\Omega_3 := \sqrt[4]{\frac{6\varepsilon p}{M_F^2 d}}$.

Method	$\bar{\delta}$	$\bar{\alpha}_k$	χ_k	$\hat{\nu}$
cFD	$\frac{\sqrt{d}M_F\nu^2}{6}$	Ω_1	Ω_2	Ω_3
cGS	$dM_F\nu^2$	$\frac{\|T_k\|}{(\|T_k\|+4.5d)}\Omega_1$	$\frac{72\|T_k\|d+216d^2+(d+2)(d+4)(d+6)}{2(\|T_k\|+4.5d)}\Omega_2$	$\sqrt[4]{\frac{2(N+4.5d)}{72Nd+216d^2+(d+2)(d+4)(d+6)}}\Omega_3$
cSS	$M_F\nu^2$	$\frac{\|T_k\|}{(\|T_k\|+4.5d)}\Omega_1$	$\frac{72\|T_k\|+216d+d^2}{2d(\|T_k\|+4.5d)}\Omega_2$	$\sqrt[4]{\frac{2d(N+4.5d)}{72N+216d+d^2}}\Omega_3$
cRC	$\frac{\sqrt{d}M_F\nu^2}{6}$	$\frac{\|T_k\|}{d}\Omega_1$	Ω_2	Ω_3
cRS	$\frac{\sqrt{d}M_F\nu^2}{6}$	$\frac{\|T_k\|}{d}\Omega_1$	Ω_2	Ω_3

3.2 Complexity Results

We now present the iteration and sample complexity results for the central finite-difference based gradient estimation methods, providing bounds on the total number of iterations and stochastic function evaluations required to achieve an ε-accurate solution, respectively. The accuracy measure is defined as follows.

Definition 1 A random iterate x_k is said to be an ε-accurate solution if it satisfies $\|\nabla F(x_k)\|^2 \leq \varepsilon$ with probability at least $1-p$, where $p \in [0,1)$.

Next, we observe that by employing $\hat{\nu}$ values in $\bar{\delta}$ in Table 1, we can ensure that

$$\|\delta_k\| = \|g(x_k) - \nabla F(x_k)\| \leq \frac{\sqrt{\varepsilon p}}{2} \leq \frac{\sqrt{\varepsilon}}{2},$$

for each gradient estimation method. Suppose that x_k satisfies $\|g(x_k)\|^2 \leq \varepsilon'$ with $\varepsilon' := \varepsilon/4$. Then, by using the above inequality, we can guarantee that

$$\|\nabla F(x_k)\| \leq \|g(x_k)\| + \|\delta_k\| \leq \sqrt{\varepsilon},$$

where the first inequality is by the triangle inequality (i.e. $\|a+b\| \leq \|a\| + \|b\|$, for any $a, b \in \mathbb{R}^n$). Therefore, for any iteration k, if $\|g(x_k)\|^2 \leq \varepsilon'$ is satisfied, then x_k is an ε-accurate solution. This observation provides a means to upper bound the sample sizes employed at each iteration before achieving an ε-accurate solution, as stated in the following lemma.

Lemma 2 Suppose the conditions of Theorem 1 hold. For all iterations $k \in \mathbb{Z}_+$ such that $\|g(x_k)\|^2 > \varepsilon'$, we have

$$|S_k| \leq b_1 + \frac{b_2}{\varepsilon'} = b_1 + \frac{4b_2}{\varepsilon}, \tag{14}$$

where $b_1, b_2 \in \mathbb{R}$ are constants that depend on the gradient estimation method. Table 2 summarizes the order results for b_1 and b_2 values.

Proof. Without loss of generality and choosing the minimum sample size that satisfies Condition 1 at each iteration $k \in \mathbb{Z}_+$, we have

$$|S_k| = \left\lceil \frac{\mathbb{E}_{T_k}[\mathbb{E}_{\zeta_i}[\|g_{\zeta_i,T_k}(x_k) - g_{T_k}(x_k)\|^2]]}{\theta^2 \mathbb{E}_{T_k}[\|g_{T_k}(x_k)\|^2]} \right\rceil \leq 1 + \frac{\mathbb{E}_{T_k}[\mathbb{E}_{\zeta_i}[\|g_{\zeta_i,T_k}(x_k) - g_{T_k}(x_k)\|^2]]}{\theta^2 \mathbb{E}_{T_k}[\|g_{T_k}(x_k)\|^2]}$$
$$\leq 1 + \frac{\mathbb{E}_{T_k}[\mathbb{E}_{\zeta_i}[\|g_{\zeta_i,T_k}(x_k) - g_{T_k}(x_k)\|^2]]}{\theta^2 \|\mathbb{E}_{T_k}[g_{T_k}(x_k)]\|^2} \leq 1 + \frac{\mathbb{E}_{T_k}[\mathbb{E}_{\zeta_i}[\|g_{\zeta_i,T_k}(x_k) - g_{T_k}(x_k)\|^2]]}{\theta^2 \varepsilon'}$$

where the second inequality is by Jensen's inequality and the last inequality is due to $\|\mathbb{E}_{T_k}[g_{T_k}(x_k)]\|^2 = \|g(x_k)\|^2 > \varepsilon'$. Following a similar procedure in analyzing the variance terms for each gradient estimation method provided in Bollapragada et al. (2024), Lemma 3.9 and adapting them to central finite-differences, and using $\varepsilon' = \frac{\varepsilon}{4}$ completes the proof. □

Now, we present the complexity results based on these sample size bounds.

Theorem 2 (Complexity Results). Suppose the conditions for Theorem 1 hold. Then, the number of iterations (iteration complexity) and the total number of stochastic function evaluations (sample complexity) required to obtain an ε-accurate solution are given as follows.

$$K_\varepsilon = \frac{8(F(x_0) - F^*)}{\varepsilon p \bar{\alpha}} = \mathscr{O}(\varepsilon^{-1}), \quad \text{and} \quad W_\varepsilon := \sum_{k=0}^{K_\varepsilon} 2|T_k||S_k| \leq 2NK_\varepsilon \left(b_1 + \frac{4b_2}{\varepsilon}\right) = \mathscr{O}(\varepsilon^{-2}). \tag{15}$$

The results are detailed in Table 2 for each gradient estimation method.

Proof. Substituting the values for $\nu = \hat{\nu}$ given in Table 1 leads to $\frac{4\bar{\chi}}{p} = \frac{\varepsilon}{2}$. From (12), it follows that for any $K \geq K_\varepsilon = \frac{8(F(x_0)-F^*)}{\varepsilon p \bar{\alpha}}$,

$$\min_{0 \leq k \leq K-1} \|\nabla F(x_k)\|^2 \leq \frac{4(F(x_0) - F^*)}{K\bar{\alpha}p} + \frac{4\bar{\chi}}{p} \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Plugging in the values for $\bar{\alpha}$ from Table 1 and for b_1 and b_2 from Table 2 completes the proof. □

Theorem 2 establishes that the iteration complexities of the methods are $\mathscr{O}(\varepsilon^{-1})$, and the sample complexities are $\mathscr{O}(\varepsilon^{-2})$, matching the optimal worst-case complexity results for nonconvex problem settings (Bottou et al. 2018; Ghadimi and Lan 2013). Similar observations to those made in forward finite-differences methods (Bollapragada et al. 2024) are found here. The results in Table 2 highlight that the cFD method exhibits the best iteration complexity compared to other methods concerning the dependency on problem dimension d. Moreover, the complexities of cGS and cSS methods are $\frac{N+d}{N}$ times worse than that of cFD, while the complexities of cRC and cRS methods are $\frac{d}{N}$ times worse than cFD. In terms of sample complexities, cFD, cRC, and cRS methods match the optimal complexity, also in terms of the dependency on d (Ghadimi and Lan 2013). Assuming ε^{-2} is the dominating term and N is small, cGS and cSS methods exhibit complexities that are d times worse than the cFD method.

4 NUMERICAL EXPERIMENTS

In this section, we evaluate the empirical performance of the proposed methods on nonlinear least squares (NLLS) problems obtained by introducing stochastic Gaussian noise of the form $\zeta \sim \mathscr{N}(0, \sigma^2 I_p)$ with $\sigma = 10^{-3}$ to the nonlinear functions $\phi : \mathbb{R}^d \to \mathbb{R}^p$ from the CUTEr (Gould et al. 2003) collection of optimization problems. We consider two different nonlinear functions ϕ: Bdqrtic $(d = 50, p = 92)$ and

Table 2: Complexity results for various gradient estimation methods. Here, β_1 and β_2 are constants defined in Assumption B. The order results specify only dependencies on d, L_F, M_f, N, ν, p, β_1, and β_2. For K_ε and W_ε, dependencies on β_1 and β_2 are removed.

Method	b_1	b_2	K_ε	W_ε
cFD	$\mathscr{O}(\beta_1)$	$\mathscr{O}(\beta_1 M_f^2 \nu^4 d + \beta_2)$	$\mathscr{O}(\frac{L_F}{\varepsilon p})$	$\mathscr{O}(\frac{L_F d}{\varepsilon^2 p})$
cGS	$\mathscr{O}(\beta_1 d)$	$\mathscr{O}((\beta_1 + d) M_f^2 \nu^4 d^3 + \beta_2 d)$	$\mathscr{O}(\frac{L_F}{\varepsilon p}\frac{N+d}{N})$	$\mathscr{O}(\frac{L_F d(N+d)}{\varepsilon^2 p} + \frac{L_F (N+d)^2 d^2}{\varepsilon(N+d^2)})$
cSS	$\mathscr{O}(\beta_1 d)$	$\mathscr{O}((\beta_1 + d) M_f^2 \nu^4 d + \beta_2 d)$	$\mathscr{O}(\frac{L_F}{\varepsilon p}\frac{N+d}{N})$	$\mathscr{O}(\frac{L_F d(N+d)}{\varepsilon^2 p} + \frac{L_F (N+d)^2 d^2}{\varepsilon(N+d^2)})$
cRC	$\mathscr{O}(\beta_1)$	$\mathscr{O}(\beta_1 M_f^2 \nu^4 d + \beta_2)$	$\mathscr{O}(\frac{L_F}{\varepsilon p}\frac{d}{N})$	$\mathscr{O}(\frac{L_F d}{\varepsilon^2 p})$
cRS	$\mathscr{O}(\beta_1)$	$\mathscr{O}(\beta_1 M_f^2 \nu^4 d + \beta_2)$	$\mathscr{O}(\frac{L_F}{\varepsilon p}\frac{d}{N})$	$\mathscr{O}(\frac{L_F d}{\varepsilon^2 p})$

Cube $(d = 20, p = 30)$, with two different error terms: absolute error and relative error. The resulting stochastic objective functions are defined as follows

$$f_{\text{abs}}(x,\zeta) := \sum_{j=1}^{p} (\phi_j(x) + \zeta_j)^2 - \sigma^2, \quad \text{and} \quad f_{\text{rel}}(x,\zeta) := \frac{1}{1+\sigma^2} \sum_{j=1}^{p} \phi_j^2(x)(1+\zeta_j)^2.$$

We employ the following practical test to approximately satisfy the theoretical Condition 1 by approximating the population (expectation) quantities with sampled quantities in our implementation.

Test 1 (Practical Norm Test) (Bollapragada et al. 2024, Test 1)

$$\frac{Var_{\zeta_i \in S_k^v}\left(g_{\zeta_i,T_k}(x_k)\right)}{|S_k|} \le \theta^2 \|g_{S_k,T_k}(x_k)\|^2, \quad \theta > 0, \tag{16}$$

where $S_k^v \subseteq S_k$ is any subset of S_k, and $Var_{\zeta_i \in S_k^v}\left(g_{\zeta_i,T_k}(x_k)\right) = \frac{1}{|S_k^v|-1} \sum_{\zeta_i \in S_k^v} \|g_{\zeta_i,T_k}(x_k) - g_{S_k,T_k}(x_k)\|^2$.

We evaluate this test at each iteration, and if the test fails, we append the set S_k with additional samples such that S_k satisfies

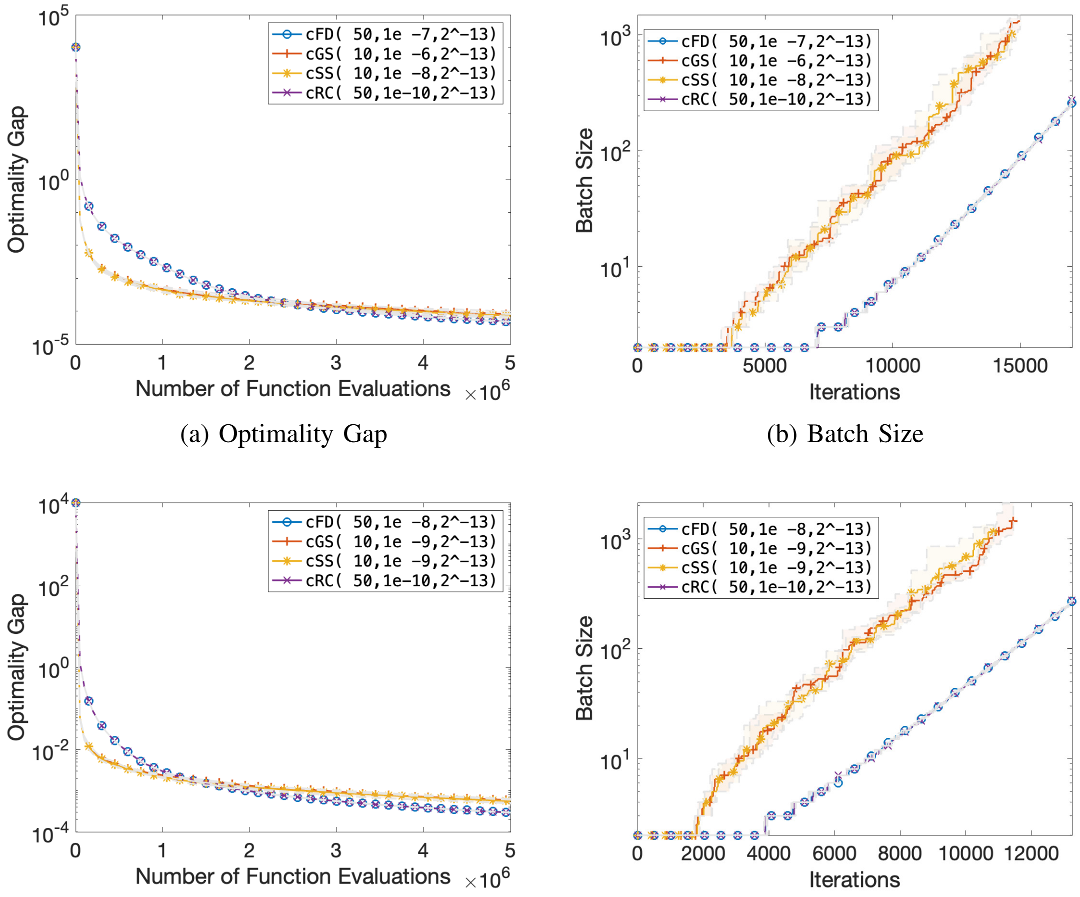
$$|S_k| = \left\lceil \frac{Var_{\zeta_i \in S_k^v}\left(g_{\zeta_i,T_k}(x_k)\right)}{\theta^2 \|g_{S_k,T_k}(x_k)\|^2} \right\rceil.$$

Although this approach involves approximations, the accuracy of these approximations improves with increasing sample sizes $|S_k|$. Moreover, this approach is more efficient than selecting a fixed, yet large, number of samples throughout the algorithm.

In our implementation, we set $\theta = 0.9$, initial sample size $|S_0| = 2$, and treat the number of sampled directions N, the sampling radius ν, and the step size α as tunable hyperparameters. We consider the worst-case performance of each combination of N, ν, and α values for each method, across three random runs. Finally, we select the best combination yielding the smallest optimality gap $(F(x_k) - F^*)$ after the methods have utilized a fixed budget of function evaluations. We conduct an additional 17 random runs for the best combinations. The legends of the figures in this section indicate the method and hyperparameters in "(method,N,ν,α)" format. We exclude the cRS method in reporting the results as its corresponding results are similar to those of the cRC method. For more detailed information about the implementation, refer to Bollapragada et al. (2024).

Figures 1 and 2 present results for the Bdqrtic and Cube functions. The solid lines represent the median performances, while the bands around the lines indicate the 35th and 65th quantiles across 20 random runs. The hyperparameters (N,ν,α) are independently tuned for each method. The tuned cRC and cFD performed similarly since the optimal N was equal to d. In Figs. 1a, 1c, 2a, and 2c, we report the optimality gap $(F(x_k) - F^*)$ with respect to the number of stochastic function evaluations. We observe that the optimal tuned number of directions (N) is smaller for smoothing methods (cGS and cSS) compared to the cRC method. Furthermore, while the smoothing methods initially exhibit superior performance, the standard finite-difference method (cFD) ultimately matches their performance as the number of function evaluations increases due to their superior accuracy in gradient estimation. In Figs. 1b, 1d, 2b, and 2b, we report the batch size or sample size ($|S_k|$) with respect to iterations. Here, the smoothing methods tend to increase the sample sizes at a faster rate than the cRC and cFD methods. We conjecture that this behavior is due to the high variance in the smoothing methods with smaller N values, necessitating larger sample sizes.

We also analyze the effect of the number of directions (N) on the empirical performance of cGS, cSS, and cRC methods in Fig. 3, which reports the optimality gap with respect to the number of function evaluations. The hyperparameters ν and α are tuned for each N and method combination. The solid lines represent the mean performance, while the bands depict the minimum and maximum values across three random runs. We observe that the behavior across different N values is problem-specific. Typically, smaller values of N lead to high variance in gradient estimation, while larger values result in high per-iteration function evaluations, rendering both inefficient. An optimal N exists that achieves the best performance for the smoothing methods.



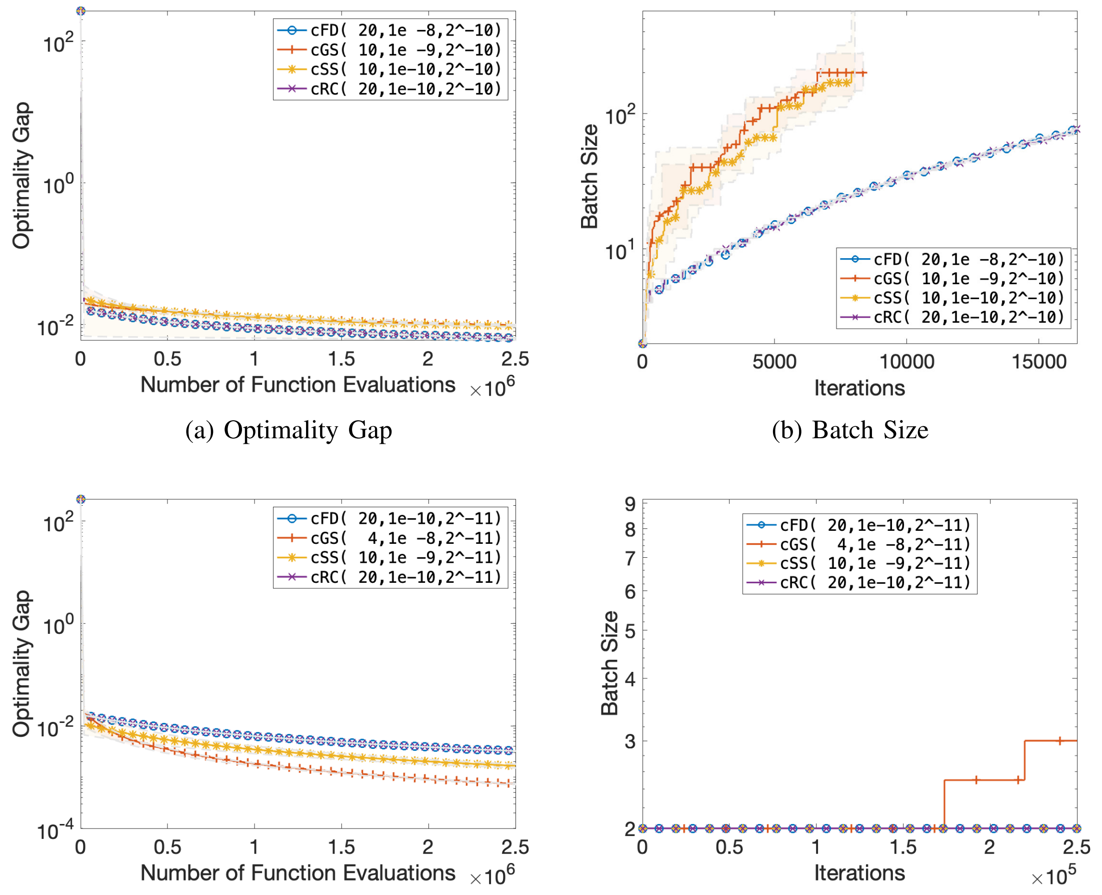
(a) Optimality Gap

(b) Batch Size

(c) Optimality Gap

(d) Batch Size

Figure 1: Performance of different gradient estimation methods using the tuned hyperparameters on the Bdqrtic function with $\sigma = 10^{-3}$. Top row: absolute error, bottom row: relative error.



(a) Optimality Gap

(b) Batch Size

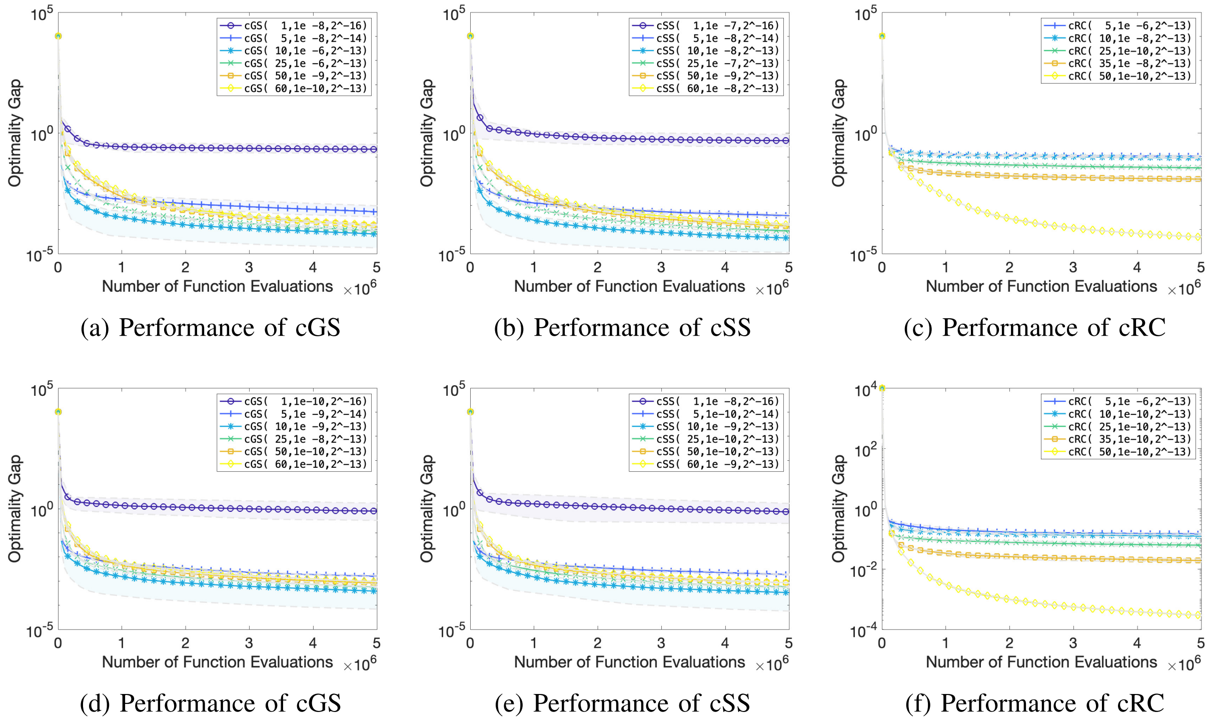
(c) Optimality Gap

(d) Batch Size

Figure 2: Performance of different gradient estimation methods using the tuned hyperparameters on the Cube function with $\sigma = 10^{-3}$. Top row: absolute error, bottom row: relative error.

5 FINAL REMARKS

Gradient estimation methods that utilize stochastic function evaluations to estimate gradients and incorporate these estimators into standard optimization methods offer scalability to large-dimensional problems and are of interest from both theoretical and practical perspectives. We introduce a unified algorithmic framework for solving stochastic optimization problems using central finite-difference based gradient estimation methods. The variance in these estimators is controlled by adaptively selecting the sample sizes employed in the stochastic approximations. We establish sublinear convergence to a neighborhood for nonconvex functions and demonstrate that this framework achieves optimal worst-case iteration and sample complexities of $\mathscr{O}(\varepsilon^{-1})$ and $\mathscr{O}(\varepsilon^{-2})$, respectively. Our numerical results on nonlinear least squares problems illustrate the effectiveness of this approach. Furthermore, this framework can be extended to explore other potentially new or hybrid variants of existing central finite-difference methods to further enhance efficiency. It can also be seamlessly integrated into more sophisticated algorithms such as quasi-Newton or accelerated methods.



(a) Performance of cGS

(b) Performance of cSS

(c) Performance of cRC

(d) Performance of cGS

(e) Performance of cSS

(f) Performance of cRC

Figure 3: The effect of number of directions N on the performance of different randomized gradient estimation methods on the Bdqrtic function with $\sigma = 10^{-3}$. Sampling radius ν and step size α are tuned for each method and N combination to achieve the best performance. Top row: absolute error, bottom row: relative error.

REFERENCES

Berahas, A. S., R. Bollapragada, and B. Zhou. 2022. “An adaptive sampling sequential quadratic programming method for equality constrained stochastic optimization”. *arXiv preprint arXiv:2206.00712*.

Berahas, A. S., R. H. Byrd, and J. Nocedal. 2019. “Derivative-free optimization of noisy functions via quasi-Newton methods”. *SIAM Journal on Optimization* 29(2):965–993.

Berahas, A. S., L. Cao, K. Choromanski, and K. Scheinberg. 2021. “A theoretical and empirical comparison of gradient approximations in derivative-free optimization”. *Foundations of Computational Mathematics*:1–54.

Bertsekas, D. 2019. *Reinforcement learning and optimal control*. Athena Scientific.

Blanchet, J., C. Cartis, M. Menickelly, and K. Scheinberg. 2019. “Convergence rate analysis of a stochastic trust-region method via supermartingales”. *INFORMS journal on optimization* 1(2):92–119.

Blum, J. R. 1954. “Multidimensional stochastic approximation methods”. *The annals of mathematical statistics*:737–744.

Bollapragada, R., R. Byrd, and J. Nocedal. 2018. “Adaptive sampling strategies for stochastic optimization”. *SIAM Journal on Optimization* 28(4):3312–3343.

Bollapragada, R., C. Karamanli, B. Keith, B. Lazarov, S. Petrides and J. Wang. 2023. “An adaptive sampling augmented Lagrangian method for stochastic optimization with deterministic constraints”. *Computers & Mathematics with Applications* 149:239–258.

Bollapragada, R., C. Karamanli, and S. M. Wild. 2024. “Derivative-Free Optimization via Adaptive Sampling Strategies”. *arXiv preprint arXiv:2404.11893*.

Bollapragada, R. and S. M. Wild. 2023. “Adaptive sampling quasi-Newton methods for zeroth-order stochastic optimization”. *Mathematical Programming Computation* 15(2):327–364.

Bottou, L., F. E. Curtis, and J. Nocedal. 2018. “Optimization methods for large-scale machine learning”. *SIAM review* 60(2):223–311.

Byrd, R. H., G. M. Chin, J. Nocedal, and Y. Wu. 2012. “Sample size selection in optimization methods for machine learning”. *Mathematical programming* 134(1):127–155.

Conn, A. R., K. Scheinberg, and L. N. Vicente. 2009. *Introduction to derivative-free optimization*. SIAM.

Flaxman, A. D., A. T. Kalai, and H. B. McMahan. 2005. “Online convex optimization in the bandit setting: gradient descent without a gradient”. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, 385–394.

Ghadimi, S. and G. Lan. 2013. “Stochastic first-and zeroth-order methods for nonconvex stochastic programming”. *SIAM journal on optimization* 23(4):2341–2368.

Gould, N. I., D. Orban, and P. L. Toint. 2003. “CUTEr and SifDec: A constrained and unconstrained testing environment, revisited”. *ACM Transactions on Mathematical Software (TOMS)* 29(4):373–394.

Kiefer, J. and J. Wolfowitz. 1952. “Stochastic estimation of the maximum of a regression function”. *The Annals of Mathematical Statistics*:462–466.

Kozak, D., S. Becker, A. Doostan, and L. Tenorio. 2021. “A stochastic subspace approach to gradient-free optimization in high dimensions”. *Computational Optimization and Applications* 79(2):339–368.

Larson, J., M. Menickelly, and S. M. Wild. 2019. “Derivative-free optimization methods”. *Acta Numerica* 28:287–404.

L’Ecuyer, P. and G. Yin. 1998. “Budget-dependent convergence rate of stochastic approximation”. *SIAM Journal on Optimization* 8(1):217–247.

Marrinan, L., U. V. Shanbhag, and F. Yousefian. 2023. “Zeroth-order Gradient and Quasi-Newton Methods for Nonsmooth Nonconvex Stochastic Optimization”. *arXiv preprint arXiv:2401.08665*.

Nesterov, Y. and V. Spokoiny. 2017. “Random gradient-free minimization of convex functions”. *Foundations of Computational Mathematics* 17(2):527–566.

Pasupathy, R., P. Glynn, S. Ghosh, and F. S. Hashemi. 2018. “On sampling rates in simulation-based recursions”. *SIAM Journal on Optimization* 28(1):45–73.

Shashaani, S., F. S. Hashemi, and R. Pasupathy. 2018. “ASTRO-DF: A class of adaptive sampling trust-region algorithms for derivative-free stochastic optimization”. *SIAM Journal on Optimization* 28(4):3145–3176.

Wright, S. J. 2015. “Coordinate descent algorithms”. *Mathematical Programming* 151(1):3–34.

AUTHOR BIOGRAPHIES

RAGHU BOLLAPRAGADA is an Assistant Professor in the Operations Research and Industrial Engineering graduate Program at the University of Texas at Austin. His research interests include nonlinear optimization, derivative-free optimization, distributed optimization and stochastic optimization. His email address is raghu.bollapragada@utexas.edu and his website is https://sites.google.com/view/raghub/home.

CEM KARAMANLI is a graduate student in the Operations Research and Industrial Engineering Program at the University of Texas at Austin. His research interests include nonlinear optimization, derivative-free optimization, distributed optimization, stochastic optimization and adaptive sampling. His email address is cem.karamanli@utexas.edu and his website is https://www.cemkaramanli.com/.

STEFAN M. WILD is the director of the Applied Mathematics and Computational Research (AMCR) Division in the Computing Sciences Area at Lawrence Berkeley National Laboratory. His research interests include developing model-based algorithms and software for challenging numerical optimization problems and automated learning. He holds editorial responsibilities for Mathematical Programming Computation, INFORMS Journal on Computing, Data Science in Science, and the SIAM Review. His email address is wild@lbl.gov and his website is https://wildsm.github.io/.