Proceedings of the 2020 Winter Simulation Conference
K.-H. Bae, B. Feng, S. Kim, S. Lazarova-Molnar, Z. Zheng, T. Roeder, and R. Thiesing, eds.

UNBIASED GRADIENT SIMULATION FOR ZEROTH-ORDER OPTIMIZATION

Guanting Chen

Institute for Computational and Mathematical Engineering
Stanford University
475 Via Ortega,
Stanford, CA 94305, USA

2020 Winter Simulation Conference (WSC) | 978-1-7281-9499-8/20/$31.00 ©2020 IEEE | DOI: 10.1109/WSC48552.2020.9384045

ABSTRACT

We apply the Multi-Level Monte Carlo technique to get an unbiased estimator for the gradient of an optimization function. This procedure requires four exact or noisy function evaluations and produces an unbiased estimator for the gradient at one point. We apply this estimator to a non-convex stochastic programming problem. Under mild assumptions, our algorithm achieves a complexity bound independent of the dimension, compared with the typical one that grows linearly with the dimension.

1 INTRODUCTION

In the field of optimization and machine learning, the problem of “Zeroth-order Optimization” or “Bandit Optimization”, both offline and online, has generated considerable interest in the operations research and learning communities. Such techniques are useful when explicit gradient calculation is expensive or infeasible. For instance, it could take days to get one noisy evaluation of a stochastic system, hence it becomes computationally expensive to get the full information of a function; sometimes the output is generated from a “black box” algorithm, which makes it challenging to get an accurate estimate of the gradient. Such examples occur in areas related to simulation optimization, distributed learning, parameter optimization, etc. Interested readers should refer to Conn et al. (2009) for the relevant background.

In zeroth-order optimization and bandit optimization, a natural approach is to approximate the gradient at a point by evaluating the function at one or several nearby points. The first relevant technique emerged in the 1950s (Robbins and Monro 1951; Kiefer and Wolfowitz 1952) and is known as stochastic approximation (SA). This approach mimics the simplest gradient descent using approximated or noisy gradient information, and has been explored in more detail and in different contexts (Spall 1992; Spall 1997). Recently, there has been a resurgence of interest in this topic (Flaxman et al. 2004; Agarwal et al. 2010; Duchi et al. 2015), where the authors use one or multiple evaluations of the zeroth-order information to generate improved complexity bounds.

It is well-known that when first-order information is available, algorithms leveraging the first-order information will have better complexity bounds than ones using the zeroth-order gradient approximation (Ghadimi and Lan 2013; Nesterov and Spokoiny 2017). This is because the zeroth-order approximation often uses finite difference methods, which obtain a closes approximation, but almost never be equal to the true gradient (the difference between the estimation and the true gradient is referred to as bias). However, the first-order information has expectation equal to the true gradient (it has no bias), despite that it often comes with random noise.

That being said, our motivation in this paper is to improve the zeroth-order method by eliminating the bias in such a way that it could then be used as a first-order method. This paper provides an unbiased estimator of the gradient that is model free and easy to calculate. Under the assumption that the function’s gradient is Lipschitz, our estimator takes function evaluations at four different locations and returns an unbiased estimator of the gradient. This type of estimator could be used in both online and offline optimization problems, whether they be deterministic or stochastic.

To illustrate this, by combining our estimator with the random stochastic gradient (RSG) method introduced by Ghadimi and Lan (2013), we develop an algorithm for non-convex stochastic optimization problems and achieve a $O(1/\varepsilon^2)$ complexity bound for $\mathbb{E}||\nabla f(x_R)||^2 \le \varepsilon$, compared with the $O(d/\varepsilon^2)$ complexity bound in the current literature.

Our contributions are as follows.

- We bridge the gap between zeroth-order methods and first-order methods in Stochastic Optimization and Bandit Convex Optimization. We find that at if one can query four function values, zeroth-order methods are the same as first-order methods. To the best of our knowledge, this is the first unbiased estimator of the gradient using only zeroth-order information.
- Our estimator can be used to improve many existing algorithms that assume a first-order oracle, hence the existing algorithm can maintain the same order of the complexity bound without assuming the first order oracle (need to assume four function evaluations instead). In this paper we use our estimator to improve the zeroth-order RSG algorithm such that it has the same complexity bound as the first-order RSG algorithm.

This paper is organized as follows. We conduct the literature review in section 2; we state the formulation of the gradient estimator in section 3; we apply the estimator to a smooth non-convex stochastic programming problem in Section 4; we prove the theorems in the appendix.

2 LITERATURE REVIEW

The one-point estimate technique, which approximates the gradient with one function evaluation, can be found in Spall (1997) and Flaxman et al. (2004). Let $f : \mathbb{R}^d \to \mathbb{R}$, by choosing a fixed x and taking a random unit vector u and small $\delta > 0$, we know

$$\nabla f(x) \approx \mathbb{E}\left[d\frac{f(x+\delta u)}{\delta}u\right].$$

Henceforth, by observing a random realization of $f(x+\delta u)$, one can get a relatively accurate estimate of the expected gradient. Such a method works well with Bandit Convex Optimization because gradient descent is a common and powerful algorithm in this problem. Agarwal et al. (2010) take this one step further by assuming that, at each round, one can observe a two-point evaluation as opposed to a one-point evaluation, and taking the gradient estimator to be

$$\frac{d}{2\delta}(f(x+\delta u) - f(x-\delta u))u,$$

they can achieve an improved regret bound. Duchi et al. (2015) apply the similar two-point estimation idea into stochastic convex optimization and obtain an improved convergence rate. For other types of one-point or two-point gradient approximation techniques and the corresponding regret or complexity bound, we refer the reader to Shamir (2013), Ghadimi and Lan (2013), Hazan and Levy (), and Nesterov and Spokoiny (2017).

Such gradient estimation techniques are often used in the following problems:

- Stochastic Optimization: In this problem the goal is to $\min_{x\in\mathbb{D}} f(x,\Xi)$ where

$$f(x,\Xi) := \mathbb{E}_P[F(x,\xi)] = \int_\Xi F(x;\xi)dP(\xi),$$

$\mathbb{D} \subseteq \mathbb{R}^d$ is a set and its property is problem specific, and P is a probability distribution over the space Ξ. For more details we refer the reader to Shamir (2013), Ghadimi and Lan (2013), Duchi et al. (2015), Shamir (2017), and Nesterov and Spokoiny (2017).

- Bandit Convex Optimization: This problem can be understood as a game between an agent and an adversary. At round t, the agent begins by choosing a point x_t in a convex set $\mathscr{K}$, then the adversary, knowing the x_t, chooses a convex loss function $f_t : \mathscr{K} \to \mathbb{R}$ and the agent observes the loss function $f(x_t)$. The agent's goal is to minimize the regret R_T which is defined as

$$R_T = \sum_{t=1}^{T} f_t(x_t) - \min_{z \in \mathscr{K}} \sum_{t=1}^{T} f_t(z).$$

For this line of literature we refer the reader to Flaxman et al. (2004), Agarwal et al. (2010), Hazan and Levy (), and Shamir (2017).

For the above problem, if one assumes that there is an oracle producing an unbiased estimator for the corresponding gradient, an algorithm can usually perform better (Nemirovski et al. 2009; Ghadimi and Lan 2013; Nesterov and Spokoiny 2017). This is mainly because the zeroth-order methods only give a randomized approximation to the gradient with bias (expectation is not equal to the gradient), while the first-order oracle directly produces the unbiased estimation. As a result, the SA method with an unbiased gradient estimator (Robbins-Monro type) has a canonical asymptotic convergence rate of $n^{-1/2}$, in contrast to $n^{-1/3}$ for the SA with the biased gradient estimator (Kiefer-Wolfowitz type). Ghadimi and Lan (2013) prove that in smooth non-convex optimization, with a first-order oracle producing unbiased estimators for the gradient, there exists an algorithm generating $\bar{x}$ such that $\mathbb{E}||\nabla f(\bar{x})||^2 < \varepsilon$ in $O(1/\varepsilon^2)$ iterations, whereas in the situation where only zeroth-order information is available, the iteration number has an upper bound of $O(d/\varepsilon^2)$. Similarly, Nesterov and Spokoiny (2017) develop a class of algorithms producing $\bar{x}$ such that $|\mathbb{E}[f(\bar{x})] - \min_{x \in K} f(x)| < \varepsilon$, and they prove upper complexity bounds for convex stochastic optimization under different settings. The algorithm with the first-order oracle has the upper bound $O(d/\varepsilon^2)$. When only zeroth-order information is available, the algorithm with the biased estimator has a complexity upper bound of $O(d^2/\varepsilon^2)$.

Although first-order methods have better convergence results, gradient information can be hard to obtain. If we do not have access to an oracle that provides unbiased estimators, we have to rely on other unbiased estimation techniques, such as infinitesimal perturbation analysis (IPA), score function (SF) methods (also called likelihood ratio (LR) method), and weak derivative (WD) methods. These methods are frequently used in simulation optimization, but have restrictions that often require full knowledge of the density function, or that are too model specific. We refer to L'Ecuyer (1990) and Fu (2006) for relevant background.

Our paper is also related to the literature on simulation. The major technique we use is called Multi-Level Monte Carlo (MLMC), which is used as a debiasing or variance reduction tool in the fields of applied probability and simulation. MLMC is a general approach to construct unbiased estimators based on a family of biased estimators. The main idea behind MLMC is to use a simple randomization technique to modify a sequence of approximations such that the modified approximations no longer have bias. Interested readers could see Giles (2008), Rhee and Glynn (2015), and Blanchet et al. (2019) for more application.

3 FORMULATION

We study the case where at time t, one can select a vector $\mathbf{x}_t = [x_t^{(1)}, x_t^{(2)}, x_t^{(3)}, x_t^{(4)}]$ as input and receive four corresponding observations $Y_t(\mathbf{x}_t)$ where

$$Y_t(\mathbf{x}_t) = [f(x_t^{(1)}, \xi_t), f(x_t^{(2)}, \xi_t), f(x_t^{(3)}, \xi_t), f(x_t^{(4)}, \xi_t)].$$

In other words, we require the same seed $\xi_t \in \Xi$ to generate the noise of four observations at time t. Similar settings could be found in (Duchi et al. 2015). Another relevant technique in simulation optimization, known as Common Random Numbers (CRN), use the same seed to reduce variance. Also notice that this includes the settings in optimization or adversarial learning where we receive exact instead of noisy feedback.

Throughout the paper, we denote $\mathbb{R}^d$ to be the standard d-dimensional Euclidean space, $\mathbb{D} \subseteq \mathbb{R}^d$ and $||\cdot||$ to be the l^2 norm. Let $C^n(\mathbb{D})$ be the set of functions that is n times continuously differentiable, and denote $C_L^{1,1}(\mathbb{D})$ to be the function class such that the gradient is Lipschitz with constant L:

$$||\nabla f(y) - \nabla f(x)|| \leq L||y-x||, \quad x,y \in \mathbb{D}.$$

From mean value theorem, this is equivalent to state that for any $x,y \in \mathbb{D}$,

$$||f(y) - f(x) - \nabla f(x)^T(y-x)|| \leq L||y-x||^2/2.$$

The equation above yields the following lemma.

Lemma 1 For any $y,x \in \mathbb{D}$, let $u = y-x$. There exists $C(x,y) : \mathbb{D}\times\mathbb{D} \to \mathbb{R}$ such that $|C(x,y)| \leq L/2$, and

$$f(y) = f(x) + \nabla f(x)^T u + C(x,y)||u||^2. \tag{1}$$

The rest of the section is organized as follows: we first derive our unbiased estimator in one dimension for intuition, and then move to the d-dimensional case.

3.1 One-Dimensional Gradient Estimation

For convenience, with a slight abuse of notation, let $f(\cdot)$ denote $f(\cdot,\omega)$. This is because we assume that at each round the noise is the same for function evaluations. Assume that the objective function $f \in C_L^{1,1}(\mathbb{D})$, where $\mathbb{D} \subseteq \mathbb{R}$ is the domain for the optimization problem. For any interior points $x \in \mathbb{D}$, consider a sequence $\{x_n\}_{n=1}^{\infty} \in \mathbb{D}$ such that $x_n \to x$. Then, we introduce another random variable $N \in \mathbb{N}$ with density function $P(N=i) = p_i$. For now we ignore the absolute convergence issue and focus on the intuition.

$$\begin{aligned} f'(x) &= \lim_{n\to\infty} \frac{f(x_n)-f(x)}{x_n - x} = \frac{f(x_1)-f(x)}{x_1-x} + \sum_{i=1}^{+\infty}\left(\frac{f(x_{i+1})-f(x)}{x_{i+1}-x} - \frac{f(x_i)-f(x)}{x_i-x}\right) \\ &= \frac{f(x_1)-f(x)}{x_1-x} + \sum_{i=1}^{+\infty}\left[\left(\frac{f(x_{i+1})-f(x)}{x_{i+1}-x} - \frac{f(x_i)-f(x)}{x_i-x}\right)\frac{1}{p_i}\cdot p_i\right]. \end{aligned} \tag{2}$$

Notice that (2) can be written as a expectation of random variable N. More specifically, If we define

$$U_x = \frac{f(x_1)-f(x)}{x_1-x} + \left(\frac{f(x_{N+1})-f(x)}{x_{N+1}-x} - \frac{f(x_N)-f(x)}{x_N-x}\right)\Big/ p_N, \tag{3}$$

$\mathbb{E}[U_x]$ equals the last line of (2), hence $\mathbb{E}[U_x] = f'(x)$.

We now state some theoretical results on this estimator.

Assumption 1 We have a sequence $\{x_i\}_{i\in\mathbb{N}} \in \mathbb{D}$ such that $|x_i - x| = c\delta^i$, where $c > 0$ and $\delta \in (0,1)$. N follows a geometric distribution such that $P(N=i) = p^{i-1}(1-p)$. Moreover, $p \in (0,1)$ and $\delta < p$.

Theorem 1 Under Assumption 1, for $f \in C_L^{1,1}(\mathbb{D})$, U_x in (3) is well-defined, $\mathbb{E}[U_x] = f'(x)$, and $\mathbb{E}[U_x^2]$ is finite.

We show the proof in Section 4. Notice that the constant c is useful in multi-dimensional settings and could be ignored when $d=1$. Since some properties of the one-dimensional estimator will be used later, we state a relevant proposition.

Proposition 1 Under Assumption 1, we have

$$\mathbb{E}[(U_x - f'(x))^2] \leq 2c^2L^2\delta^2 + 2\frac{c^2L^2\delta^2}{p(1-p)(1-\delta^2/p)}.$$

Remark 1 For the unbiasedness, there are indeed much less restrictions on the choice of δ and p than Assumption 1. We choose a geometric distribution for p_N and a geometric shrinking distance because it is convenient for the proof. One can choose any decaying sequence of $\{\delta_i\}_{i\in\mathbb{N}}$ so long as the infinite sum converges.

Algorithm 1 (Unbiased Estimator of $f'(x)$ for $x \in \mathbb{R}$)

---

1. Sample N on the distribution $P(N = n) = p_n$.
2. Conditioned on $N = n$, evaluate $f(x), f(x_1), f(x_n)$ and $f(x_{n+1})$, where $|x_i - x| = \delta^i$.
3. Output

$$U_x = \frac{f(x_1) - f(x)}{x_1 - x} + \frac{1}{p_n}\left(\frac{f(x_{n+1}) - f(x)}{x_{n+1} - x} - \frac{f(x_n) - f(x)}{x_n - x}\right).$$

3.2 Multi-Dimensional Gradient Estimation

Based on the one-dimensional result, there is an obvious way to develop the gradient estimator for a multi-dimensional case: we can simply estimate each partial derivative, and finally get an unbiased estimator of the gradient using $4d$ queries.

However, this is not the most efficient way. It turns out that with four queries of the function we can still get an unbiased estimator for the gradient. The trick is to do further randomization. This technique dates back to Spall (1997) and has been widely used in recent literature (Agarwal et al. 2010; Duchi et al. 2015; Shamir 2017; Nesterov and Spokoiny 2017). Consider a random vector $u \in \mathbb{R}^d$ with mean 0 that is uniformly distributed on its support. For example, each u_i could have an independent and uniform distribution on $\{-1, 1\}$, or it could have a uniform distribution on a l_2 sphere.

Let $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d, ||x|| = 1\}$ be the unit sphere in dimension d. Our estimator chooses u sampled uniformly on $\mathbb{S}^{d-1}$. In the multivariate case, we first estimate the directional derivative at the direction of u, denoted as $f'(x,u)$. Since estimating the directional derivative is a one-dimensional problem, we can derive an estimator, denoted by $U_{x,u}$, such that $\mathbb{E}[U_{x,u}|u] = f'(x,u)$. f being differentiable implies that $f'(x,u) = \langle \nabla f(x), u\rangle$. Multiplying u and taking the expectation yields

$$\mathbb{E}[U_{x,u}u] = \mathbb{E}[u\mathbb{E}[U_{x,u}|u]] = \mathbb{E}[f'(x,u)u] = \mathbb{E}[\langle \nabla f(x), u\rangle u]. \tag{4}$$

By calculating the expectation in each entry, it is not hard to find that $d\mathbb{E}[\langle \nabla f(x), u\rangle u] = \nabla f(x)$. Henceforth, we can give our multi-dimensional gradient algorithm below.

We then state the theorem on the multi-dimensional unbiased gradient estimator.

Assumption 2 Let $\{x_i\}_{i\in\mathbb{N}} \in \mathbb{D}$ such that $x_i - x = c\delta^i u$, where $u \sim$ Uniform $(\mathbb{S}^{d-1})$, $c > 0$ and $\delta \in (0,1)$. N follows a geometric distribution such that $P(N = i) = p^{i-1}(1-p)$. Moreover, $p \in (0,1)$ and $\delta^2 < p$.

Theorem 2 Under Assumption 2, for $f \in C_L^{1,1}(\mathbb{D})$, U_x in (5) is well defined and $\mathbb{E}[dU_x \cdot u] = \nabla f(x)$. Moreover,

$$\mathrm{Var}(dU_{x,u}) < 2M||\nabla f(x)||^2 + 4c^2d^2L^2\delta^2\left(1 + \frac{1}{p(1-p)(1-\delta^2/p)}\right), \tag{6}$$

where M is a constant independent of all other variables.

Algorithm 2 Unbiased Estimator of $\nabla f(x)$ for $x \in \mathbb{R}^d$

1. Sample N on the distribution $P(N=n) = p_n$.
2. Sample $y \sim N(0,I)$, where I is a $d \times d$ identity matrix. Let $u = y/||y||$.
3. Conditioned on $N=n$, evaluate $f(x), f(x_1), f(x_n)$ and $f(x_{n+1})$, where $x_i - x = c\delta^i \cdot u$.
4. Compute
$$U_{x,u} = \frac{f(x_1)-f(x)}{||x_1-x||} + \frac{1}{p_n}\left(\frac{f(x_{n+1})-f(x)}{||x_{n+1}-x||} - \frac{f(x_n)-f(x)}{||x_n-x||}\right). \tag{5}$$
5. Output $dU_{x,u} \cdot u$.

Remark 2 One can notice that the last term above does not grow with the dimension d if δ is chosen appropriately. This is very important for getting an improved upper bound in the next section. In fact, one can develop other types of estimator like the one above. For example, u could be sampled from multivariate gaussian distribution. One just have to calculate the coefficient to make the estimator unbiased (in our case it is d), and calculate the bound for variance.

4 APPLICATION TO NONCONVEX STOCHASTIC PROGRAMMING

In this section we begin by introducing the RSG method developed in Ghadimi and Lan (2013), which could be used when the first-order or the zeroth-order information is available (we will refer to them as the first-order RSG and the zeroth-order RSG below). We apply our unbiased estimator to the scenario where only zeroth-order information is available, generating an unbiased estimator for the gradient such that we can combine with the first-order RSG method. Our estimator achieve a $O(1/\varepsilon^2)$ complexity bound compared with the $O(d/\varepsilon^2)$ bound in Ghadimi and Lan (2013).

4.1 Introduction to the First-Order RSG Method

Ghadimi and Lan (2013) develop the randomized stochastic gradient (RSG) method for solving nonconvex stochastic programming problems in the form of

$$f^* = \inf_{x \in \mathbb{R}^d} f(x),$$

where $f \in C_L^{1,1}(\mathbb{R}^d)$ and could be non-convex. They develop the RSG method for the case where noisy first-order values could be observed. They assume that the observed value, which is a stochastic gradient, is generated by a first-order oracle. More specifically, at time k, the stochastic gradient $G(x_k, \xi_k)$ takes the input x_k chosen by the user and a random seed ξ_k which is sampled from the distribution P_k. For each k, the stochastic gradient has to satisfy the following assumption

Assumption 3
$$\mathbb{E}[G(x_k,\xi_k)] = \nabla f(x_k),\ \mathbb{E}[||G(x_k,\xi_k) - \nabla f(x_k)||^2] \le \sigma^2.$$

We summarize the result and the complexity bound below.

Theorem 3 (Ghadimi and Lan 2013) Let

$$P_R(k) = P(R=k) = \frac{2\gamma_k - L\gamma_k^2}{\sum_{k=1}^N (2\gamma_k - L\gamma_k^2)},\ \gamma_k = \min\left\{\frac{1}{L}, \frac{\tilde{D}}{\sigma\sqrt{N}}\right\},\ D_f = \sqrt{2(f(x_1)-f^*)/L},$$

for some $\tilde{D} > 0$. Under Assumption 3, we have

$$\frac{1}{L}\mathbb{E}\left[||\nabla f(x_R)||^2\right] \le \frac{LD_f^2}{N} + \left(\tilde{D} + \frac{D_f^2}{\tilde{D}}\right)\frac{\sigma}{\sqrt{N}}.$$

Algorithm 3 RSG method (Ghadimi and Lan 2013).

Input: initial point x_1, iteration number N, step sizes $\{\gamma_k\}_{k\geq 1}$ and probability density function $P_R(\cdot)$ supported on $\{1,\cdots,N\}$.

1. Sample R from the density function P_R.
2. For each step $k \leq R$, the first-order oracle produce $G(x_k,\xi_k)$, set
$$x_{k+1} = x_k - \gamma_k G(x_k,\xi_k).$$
3. Output x_R

Remark 3 This theorem yields a complexity bound of $O(1/\varepsilon^2)$ to achieve $\mathbb{E}||\nabla f(X_R)||^2 \leq \varepsilon$. Notice that this complexity bound has dependence on the second moment of the gradient estimator.

4.2 Zeroth-Order Algorithms

In the second case where only zeroth-order information is available, the objective becomes

$$f^* = \inf_{x\in\mathbb{R}^d}\left\{f(x) := \int_{\Xi} F(x,\xi)dP(\xi)\right\}.$$

It is assumed that $F(\cdot,\xi)\in C_L^{1,1}(\mathbb{R}^d)$ almost surely, hence $f\in C_L^{1,1}(\mathbb{R}^d)$. For our application we have to make the following assumptions.

Assumption 4 At each round k, we can observe a vector

$$Y_k(\mathbf{x}_k) = [F(x_k^{(1)},\xi_k),F(x_k^{(2)},\xi_k),F(x_k^{(3)},\xi_k),F(x_k^{(4)},\xi_k)],$$

where $x_k^{(i)}$ is chosen by the user. Let x be the point where we want to evaluate the gradient. Define $\{x_i\}_{i\in\mathbb{N}}$ such that $x_i - x = c\delta^i u$, where $u\sim$ Uniform $(\mathbb{S}^{d-1})$, $c>0$ and $\delta\in(0,1)$. Let N follow a geometric distribution such that $P(N=i) = p^{i-1}(1-p)$. Moreover, $p\in(0,1)$ and $\delta^2 < p$.

Then we state our algorithm and the complexity result.

Proposition 2 Under Assumption 4, with the parameter being

$$P_R(k) = P(R=k) = \frac{2\gamma_k - (L+2LM)\gamma_k^2}{\sum_{k=1}^N(2\gamma_k-(L+2LM)\gamma_k^2)},\ \gamma_k = \min\left\{\frac{1}{L+2LM},\frac{\tilde{D}}{\bar{C}(c,d,L,\delta,p)\sqrt{N}}\right\},\ \tilde{D}>0,$$
$$\bar{C}(c,d,L,\delta,p) = \sqrt{4c^2d^2L^2\delta^2\left(1+\frac{1}{p(1-p)(1-\frac{\delta^2}{p})}\right)},\ D_f = \sqrt{2(f(x_1)-f^*)/L}.$$

x_R produced by Algorithm 4 have the following property

$$\mathbb{E}\left[||\nabla f(x_R)||^2\right] \leq \frac{L(L+2LM)D_f^2}{N} + \left(\tilde{D}+\frac{D_f^2}{\tilde{D}}\right)\sqrt{\frac{2c^2d^2L^4\delta^2(1+\frac{1}{p(1-p)(1-\delta^2/p)})}{N}}. \tag{7}$$

Remark 4 The way we achieve (7) is by adopting our unbiased estimator and modifying the probability distribution function $P_R(\cdot)$. If we set $c\propto 1/d$, the MLMC RSG method could achieve a complexity bound of $O(1/\varepsilon^2)$.

Algorithm 4 MLMC RSG method

Input: Initial point x_1, iteration number N, step sizes $\{\gamma_k\}_{k\geq 1}$ and probability density function $P_R(\cdot)$ supported on $\{1,\cdots,N\}$.

1. Sample R from the density function P_R. Initialize $k=1$
While $k<R$
2. Sample N on the distribution $P(N=n)=p_n$.
3. Sample $y\sim N(0,I)$ where I is a $d\times d$ matrix. Let $u=y/||y||$.
4. Set $x=x_k$, evaluate $F(x,\xi_k)$,$F(x_1,\xi_k)$, $F(x_N,\xi_k)$ and $F(x_{N+1},\xi_k)$, where $x_k-x=c\delta^i\cdot u$.
5. Set $F(\cdot)=F(\cdot,\xi_k)$ Compute

$$U_{x,u}=\frac{F(x_1)-F(x)}{||x_1-x||}+\frac{1}{p_N}\left(\frac{F(x_{n+1})-F(x)}{||x_{n+1}-x||}-\frac{F(x_n)-F(x)}{||x_n-x||}\right).$$

6. $x_{k+1}=x_k-\gamma_k dU_{x,u}\cdot u,\ \ k=k+1.$
End while
7. Output x_R

Proof. Since there is no conflict for common assumptions, we can use the setup in Ghadimi and Lan (2013). To better connect the notations, define $G(x_k,\tilde{\xi}_k)=dU_{x_k,u}u$, notice that $\tilde{\xi}_k$ could be interpreted as the seed generating u and N at step k. Denote $\delta_k=G(x_k,\tilde{\xi}_k)-\nabla f(x_k)$, from equation (2.8) of Ghadimi and Lan (2013) we have

$$f(x_{k+1})\leq f(x_k)-\left(\gamma_k-\frac{L}{2}\gamma_k^2\right)||\nabla f(x_k)||^2-(\gamma_k-L\gamma_k^2)\langle\nabla f(x_k),\delta_k\rangle+\frac{L}{2}\gamma_k^2||\delta_k||^2. \tag{8}$$

From (6) we know that

$$||\delta_k||^2<\bar{C}(c,d,L,\delta,p)^2+2M||\nabla f(x_k)||^2.$$

Then we can modify (8) to

$$f(x_{k+1})\leq f(x_k)-\left(\gamma_k-\left(\frac{L}{2}+LM\right)\gamma_k^2\right)||\nabla f(x_k)||^2-(\gamma_k-L\gamma_k^2)\langle\nabla f(x_k),\delta_k\rangle+\frac{L}{2}\gamma_k^2\bar{C}(c,d,L,\delta,p)^2.$$

Summing it up. Since $f(x_{N+1})\geq f^*$, we have

$$\sum_{k=1}^N\left(\gamma_k-\left(\frac{L}{2}+LM\right)\gamma_k^2\right)||\nabla f(x_k)||^2\leq f(x_1)-f^*-\sum_{k=1}^N(\gamma_k-L\gamma_k^2)\langle\nabla f(x_k),\delta_k\rangle+\frac{L}{2}\sum_{k=1}^N\gamma_k^2\bar{C}(c,d,L,\delta,p)^2. \tag{9}$$

Then from a similar argument in Ghadimi and Lan (2013), let R follow the distribution that

$$P(R=k)=\frac{2\gamma_k-(L+2LM)\gamma_k^2}{\sum_{k=1}^N(2\gamma_k-(L+2LM)\gamma_k^2)}.$$

Next, the expectation could be written as

$$\mathbb{E}[||\nabla f(x_R)||^2]=\frac{\sum_{k=1}^N\left(\left(2\gamma_k-(L+2LM)\gamma_k^2\right)\mathbb{E}||\nabla f(x_k)||^2\right)}{\sum_{k=1}^N(2\gamma_k-(L+2LM)\gamma_k^2)}. \tag{10}$$

Since $\mathbb{E}[(\gamma_k - L\gamma_k^2)\langle \nabla f(x_k), \delta_k\rangle] = 0$, combing (9) and (10) we obtain

$$\frac{1}{L}\mathbb{E}[||\nabla f(x_R)||^2] \leq \frac{D_f^2 + \bar{C}(c,d,L,\delta,p)^2 \sum_{k=1}^N \gamma_k^2}{\sum_{k=1}^N (2\gamma_k - (L+2LM)\gamma_k^2)}, \tag{11}$$

where D_f is defined in proposition 2. Hence if we choose

$$\gamma_k = \min\left\{\frac{1}{L+2LM}, \frac{\tilde{D}}{\bar{C}(c,d,L,\delta,p)\sqrt{N}}\right\},$$

where $\tilde{D}$ could be any strictly positive number, we get

$$\begin{aligned}
&\frac{D_f^2 + \bar{C}(c,d,L,\delta,p)^2 \sum_{k=1}^N \gamma_k^2}{\sum_{k=1}^N (2\gamma_k - (L+2LM)\gamma_k^2)} = \frac{D_f^2 + N\bar{C}(c,d,L,\delta,p)^2\gamma_1^2}{N\gamma_1(2-(L+2LM)\gamma_1)} \leq \frac{D_f^2}{N\gamma_1} + \bar{C}(c,d,L,\delta,p)^2\gamma_1 \\
&\leq \frac{D_f^2}{N}\max\left\{L+2LM, \frac{\bar{C}(c,d,L,\delta,p)\sqrt{N}}{\tilde{D}}\right\} + \frac{\tilde{D}\bar{C}(c,d,L,\delta,p)}{\sqrt{N}} \\
&\leq \frac{(L+2LM)D_f^2}{N} + \left(\tilde{D} + \frac{D_f^2}{\tilde{D}}\right)\frac{\bar{C}(c,d,L,\delta,p)}{\sqrt{N}}.
\end{aligned} \tag{12}$$

Finally, by combining (11) and (12), we arrive at the conclusion that

$$\mathbb{E}\left[||\nabla f(x_R)||^2\right] \leq \frac{L(L+2LM)D_f^2}{N} + \left(\tilde{D} + \frac{D_f^2}{\tilde{D}}\right)\sqrt{\frac{2c^2d^2L^4\delta^2(1+\frac{1}{p(1-p)(1-\delta^2/p)})}{N}}.$$

□

5 CONCLUSION

In this paper, we develop an unbiased estimator for a class of function gradients using four function evaluations. We analyze the expectation and the variance of this estimator in one and multiple dimensions. To show the theoretical value of this estimator, we adapt our estimator to the RSG algorithm and get an improved bound for zeroth-order stochastic optimization problem.

A PROOFS

Proof of Theorem 1. Firstly we have to justify the infinite sum. Using (1), we have

$$\sum_{i=1}^{+\infty}\left|\frac{f(x_{i+1})-f(x)}{x_{i+1}-x} - \frac{f(x_i)-f(x)}{x_i - x}\right| = \sum_{i=1}^{+\infty}|C(x_{i+1},x)(x_{i+1}-x) - C(x_i,x)(x_i-x)| \leq c\sum_{i=1}^{+\infty} L\delta^i. \tag{13}$$

The condition for the above infinite sum to converge is $\delta \in (0,1)$, hence the equation

$$\begin{aligned}
f'(x) = \lim_{n\to\infty}\frac{f(x_n)-f(x)}{x_n - x} &= \frac{f(x_1)-f(x)}{x_1-x} + \sum_{i=1}^{+\infty}\left(\frac{f(x_{i+1})-f(x)}{x_{i+1}-x} - \frac{f(x_i)-f(x)}{x_i-x}\right) \\
&= \frac{f(x_1)-f(x)}{x_1-x} + \sum_{i=1}^{+\infty}\left[\left(\frac{f(x_{i+1})-f(x)}{x_{i+1}-x} - \frac{f(x_i)-f(x)}{x_i-x}\right)\frac{p_i}{p_i}\right]
\end{aligned}$$

holds. Consequently, the estimator U_x in (3) is well defined. The unbiasedness holds trivially by writing down the expectation directly, canceling all the p_i term.

It remains to show the property such as boundedness and finite variance. We begin by proving the boundedness first. Notice that it suffices to show the boundedness of $\left|\frac{1}{p_N}\left(\frac{f(x_{N+1})-f(x)}{x_{N+1}-x}-\frac{f(x_N)-f(x)}{x_N-x}\right)\right|$.

Reusing (13) yields

$$\left|\frac{1}{p_N}\left(\frac{f(x_{N+1})-f(x)}{x_{N+1}-x}-\frac{f(x_N)-f(x)}{x_N-x}\right)\right| \leq cL\delta^N/p_N.$$

Hence taking $p_N=(1-p)p^{N-1}$ such that $p>\delta$ will ensure boundedness property.

Although boundedness implies finite second moment, the actual condition of finite second moment on δ, p is weaker. It suffices to show that

$$\mathbb{E}\left[\left|\frac{1}{p_N}\left(\frac{f(x_{N+1})-f(x)}{x_{N+1}-x}-\frac{f(x_N)-f(x)}{x_N-x}\right)\right|^2\right] < +\infty.$$

Using (13) again, under the condition $\delta^2<p<1$, we have

$$\begin{aligned}\mathbb{E}\left[\left|\frac{1}{p_N}\left(\frac{f(x_{N+1})-f(x)}{x_{N+1}-x}-\frac{f(x_N)-f(x)}{x_N-x}\right)\right|^2\right] &\leq c^2L^2\mathbb{E}\left[\delta^{2N}/((1-p)^2p^{2N})\right] \\ =\frac{c^2L^2}{1-p}\sum_{i=1}^{\infty}\delta^{2i}/p^i = \frac{c^2L^2\delta^2}{p(1-p)(1-\delta^2/p)}.\end{aligned} \tag{14}$$

Hence choosing $\delta^2<p<1$ would ensure a finite second moment. □

Proof of Proposition 1.

$$\mathbb{E}[(U_x-f'(x))^2] = \mathbb{E}\left[\left(\frac{f(x_1)-f(x)}{x_1-x}-f'(x)+\frac{1}{p_N}\left(\frac{f(x_{N+1})-f(x)}{x_{N+1}-x}-\frac{f(x_N)-f(x)}{x_N-x}\right)\right)^2\right].$$

Applying the inequality $(a+b)^2 \leq 2a^2+2b^2$ yields

$$\mathbb{E}[(U_x-f'(x))^2] \leq 2\mathbb{E}\left[\left(\frac{f(x_1)-f(x)}{x_1-x}-f'(x)\right)^2\right]+2\mathbb{E}\left[\left(\frac{1}{p_N}\left(\frac{f(x_{N+1})-f(x)}{x_{N+1}-x}-\frac{f(x_N)-f(x)}{x_N-x}\right)\right)^2\right]. \tag{15}$$

The second term has a bound from (14), and for the first term, Taylor's theorem ensure the existence of $\bar{x}\in(-c\delta, c\delta)$ such that $(f(x_1)-f(x))/(x_1-x)=f'(\bar{x})$. Then from $f\in C_L^{1,1}$ we know

$$\mathbb{E}\left[\left(\frac{f(x_1)-f(x)}{x_1-x}-f'(x)\right)^2\right] = \mathbb{E}\left[\left(f'(\bar{x})-f'(x)\right)^2\right] \leq L^2\mathbb{E}\left[(\bar{x}-x)^2\right] \leq c^2L^2\delta^2. \tag{16}$$

Combining (14) (15), and (16) we conclude with

$$\mathbb{E}[(U_x-f'(x))^2] \leq 2c^2L^2\delta^2+\frac{2c^2L^2\delta^2}{p(1-p)(1-\delta^2/p)}.$$

Proof of Theorem 2: We first prove the unbiasedness result. From (4) we know that

$$\mathbb{E}[dU_{x,u}u] = d\mathbb{E}[\langle \nabla f(x), u\rangle u].$$

Hence it suffices to prove that $\mathbb{E}[u_i^2] = 1/d$. This could be done by a symmetrical argument. Define $Y_i = u_i^2/||u||^2$, by observing $\mathbb{E}[Y_i] = \mathbb{E}[Y_j]$ for $i \neq j$ and the fact that $\mathbb{E}[\sum_{i=1}^d Y_i] = 1$ we can conclude $\mathbb{E}[u_i^2] = 1/d$, hence finish the unbiasedness proof.

For the bound on variance, using triangle inequality we obtain

$$\mathbb{E}\left[||dU_{x,u}u - \nabla f(x)||^2\right] \leq \mathbb{E}\left[||dU_{x,u}u - d\langle \nabla f(x), u\rangle u||^2\right] + \mathbb{E}\left[||d\langle \nabla f(x), u\rangle u - \nabla f(x)||^2\right]. \tag{17}$$

Then we begin to analyze those two terms separately. Use the parametrization $\theta_{x,u}(t) = f(x + t \cdot u)$, and then we can rewrite $U_{x,u}$ as

$$U_{x,u} = \frac{\theta_{x,u}(c\delta) - \theta_{x,u}(0)}{c\delta} + \frac{1}{p_N}\left(\frac{\theta_{x,u}(c\delta^{N+1}) - \theta_{x,u}(0)}{c\delta^{N+1}} - \frac{\theta_{x,u}(c\delta^N) - \theta_{x,u}(0)}{c\delta^N}\right).$$

Therefore,

$$\mathbb{E}\left[||dU_{x,u}u - d\langle \nabla f(x), u\rangle u||^2\right] = d^2\mathbb{E}\left[\left(U_{x,u} - \langle \nabla f(x), u\rangle\right)^2 ||u||^2\right] = d^2\mathbb{E}\left[||u||^2\,\mathbb{E}\left[\left(U_{x,u} - \langle \nabla f(x), u\rangle\right)^2 |u\right]\right]$$
$$= d^2\mathbb{E}\left[\mathbb{E}\left[\left(U_{x,u} - \theta'_{x,u}(0)\right)^2 |u\right]\right].$$

Notice that for every x, u,

$$|\theta'_{x,u}(t_1) - \theta'_{x,u}(t_2)| = |\langle \nabla f(x + t_1 \cdot u) - \nabla f(x + t_2 \cdot u), u\rangle| \leq ||\langle \nabla f(x + t_1 \cdot u) - \nabla f(x + t_2 \cdot u)||$$
$$\leq L||(t_1 - t_2) \cdot u|| = L|t_1 - t_2|.$$

Hence we know that $\theta_{x,u}(\cdot) \in C_L^{1,1}$ uniformly for x and u. Then with Assumption 1 and the condition $\delta^2 < p$, we know that uniformly

$$\mathbb{E}\left[\left(U_{x,u} - \theta'_{x,u}(0)\right)^2 |u\right] \leq 2c^2L^2\delta^2 + \frac{2c^2L^2\delta^2}{p(1-p)(1-\delta^2/p)}.$$

Therefore

$$\mathbb{E}\left[||dU_{x,u}u - d\langle \nabla f(x), u\rangle u||^2\right] \leq 2c^2d^2L^2\delta^2 + \frac{2c^2d^2L^2\delta^2}{p(1-p)(1-\delta^2/p)}. \tag{18}$$

Next, consider the bound of the second term in (17). Denote $f'_i(x)$ to be the i-th entry of $\nabla f(x)$.

$$\mathbb{E}\left[||d\langle \nabla f(x), u\rangle u - \nabla f(x)||^2\right] = \sum_{i=1}^d f'_i(x)^2\mathbb{E}\left[\left((du_i^2 - 1)\right)^2\right].$$

Finally, we have to give a bound on $\mathbb{E}\left[\left((du_i^2 - 1)\right)^2\right]$. From Spruill et al. (2007), the density function for u_i is

$$f_d(x) = \frac{\Gamma(d/2)}{\sqrt{\pi}\Gamma((d-1)/2)}(1 - x^2)^{(d-3)/2} I_{\{-1,1\}}(x),$$

where $I_A(\cdot)$ denote the indicator function on set A. By direct calculation $\mathbb{E}[u_i^4] = 3\Gamma(d/2)/4\Gamma(d/2+2)$. By the fact that $\Gamma(d+2)/\Gamma(d) = (d+2)(d+1)$ for $d \in \mathbb{N}$ and the well-known asymptotic limit

$$\lim_{d\to\infty} \frac{\Gamma(d+\alpha)}{\Gamma(d)d^\alpha} = 1,$$

we conclude that there exist $M > 0$ such that for any $d > 0$ and $u \in \text{Uniform}(\mathbb{S}^{d-1})$, $\mathbb{E}(d^2 u_i^2) \leq M$. Henceforth, we have

$$\mathbb{E}\left[||d\langle \nabla f(x), u\rangle u - \nabla f(x)||^2\right] = \sum_{i=1}^{d} f_i'(x)^2 \mathbb{E}\left[\left((du_i^2 - 1)\right)^2\right] \leq M||\nabla f(x)||^2. \tag{19}$$

To sum up, combining (17), (18), and (19) we have the variance bound being

$$\mathbb{E}\left[||dU_{x,u}u - \nabla f(x)||^2\right] \leq 4c^2 d^2 L^2 \delta^2 \left(1 + \frac{1}{p(1-p)(1-\delta^2/p)}\right) + 2M||\nabla f(x)||^2.$$

□

REFERENCES

Agarwal, A., O. Dekel, and L. Xiao. 2010. "Optimal Algorithms for Online Convex Optimization with Multi-Point Bandit Feedback.". In *Proceedings of the 23rd Annual Conference on Learning Theory*, Volume 23, 28–40.

Blanchet, J. H., P. W. Glynn, and Y. Pei. 2019. "Unbiased Multilevel Monte Carlo: Stochastic Optimization, Steady-state Simulation, Quantiles, and Other Applications". *https://arxiv.org/abs/1904.09929, accessed 16th February 2020.*.

Conn, A. R., K. Scheinberg, and L. N. Vicente. 2009. *Introduction to Derivative-free Optimization*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Duchi, J. C., M. I. Jordan, M. J. Wainwright, and A. Wibisono. 2015. "Optimal Rates for Zero-order Convex Optimization: The Power of Two Function Evaluations". *IEEE Transactions on Information Theory* 61(5):2788–2806.

Flaxman, A. D., A. T. Kalai, and H. B. McMahan. 2004. "Online Convex Optimization in the Bandit Setting: Gradient Descent Without a Gradient". *https://arxiv.org/abs/cs/0408007, accessed 16th February 2020.*.

Fu, M. C. 2006. "Gradient Estimation". In *Handbooks in Operations Research and Management Science*, edited by S. Henderson and B. Nelson, Volume 13, Chapter 19, 575–616. Elsevier.

Ghadimi, S., and G. Lan. 2013. "Stochastic First-and Zeroth-order Methods for Nonconvex Stochastic Programming". *SIAM Journal on Optimization* 23(4):2341–2368.

Giles, M. B. 2008. "Multilevel Monte Carlo Path Simulation". *Operations Research* 56(3):607–617.

Hazan, E., and K. Levy. "Bandit Convex Optimization: Towards Tight Bounds". In *Advances in Neural Information Processing Systems*, Volume 27, 784–792.

Kiefer, J., and J. Wolfowitz. 1952. "Stochastic Estimation of the Maximum of a Regression Function". *The Annals of Mathematical Statistics* 23(3):462–466.

L'Ecuyer, P. 1990. "A Unified View of the IPA, SF, and LR Gradient Estimation Techniques". *Management Science* 36(11):1364–1383.

Nemirovski, A., A. Juditsky, G. Lan, and A. Shapiro. 2009. "Robust Stochastic Approximation Approach to Stochastic Programming". *SIAM Journal on optimization* 19(4):1574–1609.

Nesterov, Y., and V. Spokoiny. 2017. "Random Gradient-free Minimization of Convex Functions". *Foundations of Computational Mathematics* 17(2):527–566.

Rhee, C.-h., and P. W. Glynn. 2015. "Unbiased Estimation with Square Root Convergence for SDE Models". *Operations Research* 63(5):1026–1043.

Robbins, H., and S. Monro. 1951. "A Stochastic Approximation Method". *The Annals of Mathematical Statistics*:400–407.

Shamir, O. 2013. "On the Complexity of Bandit and Derivative-free Stochastic Convex Optimization". *https://arxiv.org/abs/1209.2388, accessed 16th February 2020.*.

Shamir, O. 2017. "An Optimal Algorithm for Bandit and Zero-order Convex Optimization with Two-point Feedback". *The Journal of Machine Learning Research* 18(1):1703–1713.

Spall, J. C. 1992. "Multivariate Stochastic Approximation Using a Simultaneous Perturbation Gradient Approximation". *IEEE Transactions on Automatic Control* 37(3):332–341.

Spall, J. C. 1997. "A One-measurement Form of Simultaneous Perturbation Stochastic Approximation". *Automatica* 33(1):109–112.

Spruill, M. et al. 2007. "Asymptotic Distribution of Coordinates on High Dimensional Spheres". *Electronic Communications in Probability* 12:234–247.

AUTHOR BIOGRAPHIES

GUANTING CHEN is a Ph.D. candidate at the Institute for Computational and Mathematical Engineering at Stanford University. He graduated with B.S. degrees in Mathematics from the University of Michigan, Ann Arbor. His research interests include applied probability, stochastic modeling and operations research. His email address is guanting@stanford.edu.