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.

DISTRIBUTIONALLY CONSTRAINED STOCHASTIC GRADIENT ESTIMATION USING NOISY FUNCTION EVALUATIONS

Henry Lam

Department of Industrial Engineering &
Operations Research
Columbia University
500 W. 120th Street
New York, NY 10027 USA

Junhui Zhang

Department of Applied Physics &
Applied Mathematics
Columbia University
500 W. 120th Street
New York, NY 10027 USA

ABSTRACT

We consider gradient estimation with only noisy function evaluation, where the function can only be evaluated at values lying within a probability simplex. We are interested in obtaining gradient estimators where each (pair of) data collection or simulation run applies simultaneously to all directions at once. Our problem is motivated from the use of stochastic approximation in distributionally robust simulation analysis, which involves solving for worst-case input distributions in a black-box simulation model. In this context, conventional gradient schemes such as simultaneous perturbation face challenges as the required moment conditions that allow the "canceling" of higher-order error terms cannot be satisfied without violating the simplex constraints. We investigate a new set of required conditions on the probability distribution that governs the perturbation, which leads us to a class of implementable gradient estimators using Dirichlet mixtures. We study the statistical properties of these estimators and demonstrate their effectiveness with numerical results.

1 INTRODUCTION

We study stochastic gradient estimation with only noisy function evaluation, where the function can only be evaluated at values lying within a probability simplex. More precisely, given a real-valued function $Z(\cdot)$ that defines on an n-dimensional set of probability weights, we want to compute its (properly defined) gradient ∇Z, by using only noisy observations of Z at suitably chosen points. Our particular focus is on schemes that can simultaneously estimate all components of the gradient with only a single or a pair of observations on Z.

Gradient estimation with only noisy function evaluation arises ubiquitously in simulation optimization, when one uses stochastic approximation (SA) or gradient descent and the simulation model is only accessible as a "black-box" (e.g., Glasserman 2013; Asmussen and Glynn 2007; Fu 2006; L'Ecuyer 1991). In derivative-free optimization, this is also known as the zeroth-order gradient oracle (Ghadimi and Lan 2013). The classical Kiefer-Wolfowitz (KW) scheme (Kushner and Yin 2003) is applicable precisely in the above optimization setting. At each iteration, KW uses a pair of noisy evaluation to compute a finite difference that approximates the partial derivative at each dimension, and convergence is provably attained by suitably controlling the perturbation size and iteration step size. The number of evaluations at each iteration required by this scheme, however, grows with the problem dimension. To address this, schemes that allow a simultaneous estimation of derivatives at all dimension have been investigated. These schemes randomly generate a perturbation vector according to some well-chosen distribution, and evaluate the function at the realized perturbation. Then, through multiplying by a factor related to the generating distribution (and possibly subtracting another evaluation at the original point), an estimate of the entire gradient is obtained.

Prominent examples of such approaches include the simultaneous perturbation stochastic approximation (SPSA) (Spall 1992; Spall 1997) that uses distributions with enough masses at zero, Gaussian smoothing (Nesterov and Spokoiny 2017), and uniform sampling (Flaxman et al. 2004) to generate the perturbation vector.

Our investigation is motivated from the use of randomized gradient estimation schemes described above in *distributionally robust simulation analysis*, a recent line of work to address input uncertainty problems. The latter refers to the estimation of confidence bounds or output variability for the performance measure of interest, when the input distributions that feed into a simulation model are corrupted by uncertainty or statistical noises. In the presence of input data, such problems can be handled by using various statistical methods including bootstrap resampling (Barton and Schruben 2001; Barton et al. 2014), delta method (Cheng and Holland 2004), empirical likelihood (Lam and Qian 2016) and Bayesian approaches (Chick 2001; Zouaoui and Wilson 2003). Distributionally robust simulation analysis targets especially at high-uncertainty, nonparametric situations where the modeler may only have minimal information on the input distributions, such as moment ranges based on expert opinions or with a small amount of data (Glasserman and Xu 2014; Ghosh and Lam 2019; Ghosh and Lam 2015; Lam 2016). The method has also been proposed to address uncertainty in directions beyond typical assumptions, such as dependency in the input processes (Lam 2018). In these situations, the method aims to find worst-case bounds on the performance measure under the imposed information. It does so by solving optimization problems with objective being the target performance measure and decision variables being the uncertain input distributions. The feasible region is represented by the imposed information as constraints on the input distributions, often known as an uncertainty set or ambiguity set. If this set contains the true input distribution with high confidence, then the resulting optimal value would bound the true value of the performance measure with at least the same level of confidence guarantee.

Our studied gradient estimator is used to solve the aforementioned worst-case optimization problems. The objective function of these problems, which is the performance measure of interest, can be a black-box simulation model in general where we can only observe noisy outputs given the input distribution. Computing the worst-case performance measure requires running stochastic approximation or gradient descent, and in this case thus falls into the setting of the zeroth-order gradient oracle where a randomized perturbation is required at each iteration. Note that, since the decision variable is the input distribution, it lies in the space of probability distributions, and the objective is typically undefined or incomputable when the decision variable is outside this region. Thus, in order to run the zeroth-order gradient-based iteration, we need to obtain a gradient estimator on the constrained space, which is precisely our focus in this paper. Though not primarily motivated, other potential applications of our investigation include SA approaches used in solving bilinear games and stochastic utility problems (Nemirovski et al. 2009).

A challenge in obtaining gradient estimator via random perturbation, compared to previously studied unconstrained settings, lies in the satisfaction of the distributional and moment conditions required for the random perturbation vector. These conditions ensure that the higher-order error terms in the function evaluation or finite difference under the perturbation can be sufficiently "canceled out", so that the resulting gradient estimator enjoys good bias properties. For instance, the standard version of SPSA requires the perturbation vector to have independent, mean-zero, symmetric components with finite reciprocal moments (Spall 1992). Because of the constraints in our problem, however, the perturbation vector we generate must have correlated and unsymmetric components, and with restrictive moment behaviors, thus making these previous conditions difficult to satisfy. Moreover, challenges also arise in using the idea of central finite-differencing in these schemes, as the probability constraint may, depending on the evaluation point of interest, only allow movement in either the forward or backward direction. To overcome all these challenges, in this paper we present a new set of conditions to ensure that the perturbation vector within the probability simplex leads to higher-order cancellation. We then propose a specific set of estimators using the Dirichlet mixture that satisfies these new conditions. We show how these estimators are readily

implementable by offering precise guidelines. We also demonstrate their effectiveness in both estimating gradients and in applying to SA.

2 PROBLEM SETTING AND MOTIVATION

Consider the function $Z(\cdot) : \mathscr{S} \to \mathbb{R}$, where

$$\mathscr{S} = \{\mathbf{p} = (p_1, \ldots, p_n) \in \mathbb{R}^n : p_1 + \cdots + p_n = 1, p_i \geq 0 \ \forall i = 1, \ldots, n\}$$

is an n-dimensional probability simplex. We are interested in estimating $\nabla Z(\mathbf{p})$, the gradient of Z. To be more precise, note that, because of the probability-simplex constraint, an arbitrary perturbation of $\mathbf{p}$ needed to define a gradient may shoot outside the domain of Z, and thus, when speaking of gradient, we mean a directional gradient $\nabla Z(\mathbf{p}) \in \mathbb{R}^n$ defined such that, for any perturbation of $\mathbf{p} \in \mathscr{S}$ to $\mathbf{q} \in \mathscr{S}$, we have

$$Z(\mathbf{q}) - Z(\mathbf{p}) = \nabla Z(\mathbf{p})'(\mathbf{p} - \mathbf{q}) + o(\|\mathbf{p} - \mathbf{q}\|) \tag{1}$$

Equivalently, we can define each component of $\nabla Z(\mathbf{p}) = (Z_1(\mathbf{p}), \ldots, Z_n(\mathbf{p}))'$ as

$$Z_i(\mathbf{p}) = \lim_{\varepsilon \to 0_+} \frac{Z((1-\varepsilon)\mathbf{p} + \varepsilon \mathbf{e}_i) - Z(\mathbf{p})}{\varepsilon} \tag{2}$$

where $\mathbf{e}_i$ is a vector that takes value 1 at component i and 0 otherwise. That is, $Z_i(\mathbf{p})$ is defined by mixing $\mathbf{p}$ with a point mass at component i and evaluating the increment as the mixture parameter shrink to 0. It is straightforward to see that (1) and (2) coincide, up to a translation of $c\mathbb{1}$ for any constant c and $\mathbb{1}$ being the vector of 1's (note that (1) as defined is non-unique up to this translation, i.e., for any $\psi(\cdot)$ satisfying (1), $\psi(\cdot) + c\mathbb{1}$ also satisfies (1) for any c). In robust statistics, definition (2) represents the so-called influence function (Hampel et al. 2011) that measures the sensitivity with respect to changes in the data distribution, when viewing $Z(\mathbf{p})$ as a statistical functional on the data distribution $\mathbf{p}$.

Our goal is to estimate $\nabla Z(\mathbf{p})$ when given only the capability to output noisy function evaluations. That is, given any $\mathbf{q}$, we can output $\hat{Z}(\mathbf{q})$ such that $E[\hat{Z}(\mathbf{q})] = Z(\mathbf{q})$. Here we take the liberal view that $\nabla Z(\mathbf{p})$ can refer to any candidate that satisfies (1). We are particularly interested in estimating $\nabla Z(\mathbf{p})$ by randomly generating a probability vector, say δ, and mix it with the original probability vector $\mathbf{p}$. This can be thought as using a perturbation vector $\delta - \mathbf{p}$. Our scheme then follows by evaluating Z at the perturbed probability vector and "projecting" to get derivative estimates simultaneously for all dimensions.

More precisely, we consider several variants of estimator. The most basic scheme is to use

$$\frac{\hat{Z}((1-c)\mathbf{p} + c\delta)}{c} S(\mathbf{p}, \delta) \tag{3}$$

where $c > 0$ and $S(\mathbf{p}, \delta) \in \mathbb{R}^n$ are chosen by user. When R simulation budget is available, we would randomly draw $\delta^j, j = 1, \ldots, R$, evaluate (3) each time with an independent evaluation of $\hat{Z}$, and at the end output their average, i.e.,

$$\frac{1}{R} \sum_{j=1}^{R} \frac{\hat{Z}^j((1-c)\mathbf{p} + c\delta^j)}{c} S(\mathbf{p}, \delta^j) \tag{4}$$

where $\hat{Z}^j(\cdot), j = 1, \ldots, R$ denotes R independent function evaluations. We call (3) and (4) the *single-function-evaluation (SFE) estimator*.

In (3) and (4), $c > 0$ is a perturbation size that could depend on a given R. The vector $S(\mathbf{p}, \delta)$ represents a factor or "score function" that aims to cancel out higher-order terms in the associated Taylor series in order to control the bias.

Next, consider the use of two function evaluations, one at $(1-c)\mathbf{p} + c\delta$ and one at $\mathbf{p}$ itself. We output

$$\frac{\hat{Z}^1((1-c)\mathbf{p} + c\delta) - \hat{Z}^2(\mathbf{p})}{c} S(\mathbf{p}, \delta) \tag{5}$$

where $\hat{Z}^1(\cdot)$ and $\hat{Z}^2(\cdot)$ represent two independent evaluations, and $c > 0$ and $S(\mathbf{p}, \delta) \in \mathbb{R}^n$ are chosen by user like before. Analogously, when $2R$ simulation budget is available, we would use

$$\frac{1}{R}\sum_{j=1}^{R} \frac{\hat{Z}^{2j-1}((1-c)\mathbf{p}+c\delta^j) - \hat{Z}^{2j}(\mathbf{p})}{c} S(\mathbf{p}, \delta^j) \tag{6}$$

where $\hat{Z}^{2j-1}(\cdot), \hat{Z}^{2j}(\cdot), j = 1, \ldots, R$ are $2R$ independent function evaluations. We call (5) and (6) the *forward-function-evaluation (FFE) estimator*.

Lastly, we also consider a variant similar to central finite-differencing. Consider using two function evaluations, one at $(1-c)\mathbf{p}+c\delta$ and one at $(1+c)\mathbf{p}-c\delta$. We output

$$\frac{\hat{Z}^1((1-c)\mathbf{p}+c\delta) - \hat{Z}^2((1+c)\mathbf{p}-c\delta)}{2c} S(\mathbf{p}, \delta) \tag{7}$$

where $\hat{Z}^1(\cdot)$ and $\hat{Z}^2(\cdot)$ again represent two independent evaluations. Like FFE, when $2R$ simulation budget is available, we would use

$$\frac{1}{R}\sum_{j=1}^{R} \frac{\hat{Z}^{2j-1}((1-c)\mathbf{p}+c\delta^j) - \hat{Z}^{2j}((1+c)\mathbf{p}-c\delta^j)}{2c} S(\mathbf{p}, \delta^j) \tag{8}$$

where $\hat{Z}^{2j-1}(\cdot), \hat{Z}^{2j}(\cdot), j = 1, \ldots, R$ are $2R$ independent function evaluations. We call (7) and (8) the *central-function-evaluation (CFE) estimator*.

Let us investigate and compare the statistical errors of the above estimators in estimating the gradient. In doing so we will also highlight the involved challenges. Typically, all the estimators above have variances that scale in order $1/(Rc^2)$, as the c appears in the denominator of these estimators. To understand the bias, supposing Z is sufficiently smooth, consider the Taylor series

$$Z((1-c)\mathbf{p}+c\delta) = Z(\mathbf{p}) + c\nabla Z(\mathbf{p})'(\delta-\mathbf{p}) + \frac{c^2}{2}(\delta-\mathbf{p})'\nabla^2 Z(\mathbf{p})(\delta-\mathbf{p}) + O(c^3) \tag{9}$$

where $\nabla^2 Z(\mathbf{p})$ is a properly defined Hessian. Now, since δ and the function evaluations are independent, we can write the expectation of SFE as

$$\begin{aligned} & E\left[\frac{Z((1-c)\mathbf{p}+c\delta)}{c} S(\mathbf{p}, \delta)\right] \\ = \; & E\left[\frac{1}{c}\left(Z(\mathbf{p}) + c\nabla Z(\mathbf{p})'(\delta-\mathbf{p}) + \frac{c^2}{2}(\delta-\mathbf{p})'\nabla^2 Z(\mathbf{p})(\delta-\mathbf{p}) + O(c^3)\right) S(\mathbf{p}, \delta)\right] \\ = \; & Z(\mathbf{p})\frac{E[S(\mathbf{p}, \delta)]}{c} + \nabla Z(\mathbf{p})' E[(\delta-\mathbf{p})S(\mathbf{p}, \delta)] + \frac{c}{2} tr(E[\nabla^2 Z(\mathbf{p})(\delta-\mathbf{p})(\delta-\mathbf{p})'S(\mathbf{p}, \delta)]) + O(c^2) \end{aligned} \tag{10}$$

Thus, if we can ensure that

$$E[S(\mathbf{p}, \delta)] = 0 \tag{11}$$

$$E[(\delta-\mathbf{p})S(\mathbf{p}, \delta)] = I \tag{12}$$

$$tr(E[\nabla^2 Z(\mathbf{p})(\delta-\mathbf{p})(\delta-\mathbf{p})'S(\mathbf{p}, \delta)]) = 0 \tag{13}$$

where I is the identity matrix and 0 are zero matrices each in the suitable dimension, then SFE gives rise to a gradient estimator that has bias $O(c^2)$. If, however, (13) cannot be satisfied, then the bias could be $O(c)$, which is less desirable.

It is easy to see that FFE follows a similar, though slightly different, behavior. Its expectation is

$$\begin{aligned} & E\left[\frac{Z((1-c)\mathbf{p}+c\delta)-Z(\mathbf{p})}{c}S(\mathbf{p},\delta)\right] \\ = \quad & \nabla Z(\mathbf{p})'E[(\delta-\mathbf{p})S(\mathbf{p},\delta)]+\frac{c}{2}tr(E[\nabla^2 Z(\mathbf{p})(\delta-\mathbf{p})(\delta-\mathbf{p})'S(\mathbf{p},\delta)])+O(c^2) \end{aligned}$$

Compared with (10), the term $Z(\mathbf{p})E[S(\mathbf{p},\delta)]/c$ disappears. Thus, if we can ensure that

$$E[(\delta-\mathbf{p})S(\mathbf{p},\delta)] = I \tag{14}$$

$$tr(E[\nabla^2 Z(\mathbf{p})(\delta-\mathbf{p})(\delta-\mathbf{p})'S(\mathbf{p},\delta)]) = 0 \tag{15}$$

then we have bias $O(c^2)$, and similarly if only (14) is enforced, then bias is $O(c)$. As we will see, (11) can be easily satisfied, and thus it may appear that SFE is preferable as it only requires one function evaluation to get one "gradient sample". However, the first term $Z(\mathbf{p})E[S(\mathbf{p},\delta)]/c$ has a c in the denominator, so with a finite simulation run this term could add significantly to the variance. This is intuitive as SFE does not resemble a finite-difference scheme, the latter capturing the differencing needed in evaluating a derivative.

Next we discuss CFE. Thanks to the central finite-differencing, the first and third-order terms in the Taylor series of $Z((1-c)\mathbf{p}+c\delta)$ and $Z((1+c)\mathbf{p}-c\delta)$ cancel out, so that

$$E\left[\frac{Z((1-c)\mathbf{p}+c\delta)-Z((1+c)\mathbf{p}-c\delta)}{2c}S(\mathbf{p},\delta)\right] = \nabla Z(\mathbf{p})'E[(\delta-\mathbf{p})S(\mathbf{p},\delta)]+O(c^2)$$

Thus, to ensure a bias $O(c^2)$, we only require

$$E[(\delta-\mathbf{p})S(\mathbf{p},\delta)] = I$$

This appears to be desirable. However, note that $(1+c)\mathbf{p}-c\delta$ is not a proper mixture and may not correspond to a probability distribution.

Thus, taking into account the variance and the implementability, FFE appears the best one among the three candidates. We will see that (11) and (12) are quite easy to enforce. In this sense, CFE can result in $O(c^2)$ bias, but it is not implementable generally due to the probability simplex constraint. SFE and FFE, on the other hand, can also ensure bias $O(c)$ easily, and FFE has significantly better variance with a price of needing twice as many function evaluations.

The main challenge, however, is that without the central finite-differencing trick, it is not obvious how to lift FFE and SFE to a bias $O(c^2)$. This would be the main investigation in the next section.

To close this section, we describe how the above estimators are applied. They are primarily motivated from distributionally robust simulation analysis. Here, think of $Z(P)$ as a performance measure that depends on some input distribution P. For instance, P could denote the interarrival or service times in a queueing system, and $Z(P)$ represents some output performance measure such as the expected wait time. When P is uncertain but some information is available, we can consider computing the worst case bounds for the performance measure given by

$$\min_{P\in\mathscr{U}} Z(P) \quad \text{and} \quad \max_{P\in\mathscr{U}} Z(P) \tag{16}$$

where $\mathscr{U}$ is the uncertainty set that contains information on P. For instance, $\mathscr{U}$ can denote moment and support constraints, so that $\mathscr{U} = \{P : E_P[f_l(X)] \le \mu_l, l = 1,2,...,s, supp\ P = A\}$ for some moment functions $f_l(\cdot)$'s, $\mu_l \in \mathbb{R}$, X denotes a generic random variable under distribution P, and A is a subset in the domain of P. As another example, one could also consider a neighborhood of a baseline model measured by statistical distances such as the ϕ-divergence, in which case $\mathscr{U} = \{P : d_\phi(P,P_b) \le \eta\}$ where $d_\phi(P,P_b)$ denotes the ϕ-divergence from some baseline distribution P_b. Depending on the information, these constraints or combinations of them can be used in a given context.

One motivation of our investigated gradient estimator is to iteratively solve (16) using SA when Z is accessible only via noisy function evaluation. Note that, in many simulation analysis, P can be a continuous distribution. However, one can suitably discretize the distribution by using support points generated from a heavy-tailed distribution and solving the problem on the sampled support points (see Ghosh and Lam 2019). This reduces to solving

$$\min_{\mathbf{p}\in\mathscr{U}} Z(\mathbf{p}) \quad \text{and} \quad \max_{\mathbf{p}\in\mathscr{U}} Z(\mathbf{p}) \tag{17}$$

Then, to solve say the minimization problem in (17), we can use for instance the Frank-Wolfe (FW) method which requires at each iteration solves $\min_{q\in\mathscr{U}} \hat{\nabla} Z(\mathbf{p})'(\mathbf{q}-\mathbf{p})$. Here, in view of the discussion above, we can equivalently replace the gradient $\nabla Z(\mathbf{p})$ with any $\psi(\mathbf{p}) = c_1\nabla Z(\mathbf{p}) + c_2\mathbb{1}$ for some $c_1 > 0$ and $c_2 \in \mathbb{R}$. Taking into account the error in the gradient estimation, Frank-Wolfe stochastic approximation (FWSA) (Ghosh and Lam 2019) would solve $\min_{q\in\mathscr{U}} \hat{\psi}(\mathbf{p})'(\mathbf{q}-\mathbf{p})$ where $\hat{\psi}(\mathbf{p})$ is an estimator of $\psi(\mathbf{p})$.

3 IMPLEMENTABLE GRADIENT ESTIMATORS WITH DIRICHLET MIXTURES

In view of the discussion in Section 2, we aim to get an estimator for $\psi(\mathbf{p})$, an equivalent to $\nabla Z(\mathbf{p})$, by using random perturbation that satisfies (11) to (12). We will focus mostly on SFE and FFE as they are more readily implementable, but we would also discuss briefly how CFE can also be used if one impose extra requirement on the initial solution in the associated SA scheme. In particular, we will base our choice on the use of a mixture of Dirichlet distributions, i.e.,

$$\delta = \sum_{i=1}^{K} \theta_i \Delta_i, \quad \theta_i > 0,\ \Delta_i \sim Dir(\alpha_i),\ \alpha_i = (\alpha_{i,1}, \alpha_{i,2}, ..., \alpha_{i,n})$$

where $(\theta_1, \ldots, \theta_K) \in \mathbb{R}^K$ are probability weights. Moreover, we set the score function $S(\mathbf{p}, \delta) = \gamma(\delta - \mathbf{p})$ for some $\gamma \in \mathbb{R}$.

For our development, the following facts on Dirichlet distributions will be useful. Let $\Delta_i = (\Delta_{i,1}, \Delta_{i,2}, ..., \Delta_{i,n}) \sim Dir(\alpha_{i,1}, \alpha_{i,2}, ..., \alpha_{i,n})$, and $\alpha_{i,0} = \sum_{j=1}^{n} \alpha_{i,j}$. We have

$$E(\Delta_{i,j}) = \frac{\alpha_{i,j}}{\alpha_{i,0}}$$

$$Var(\Delta_{i,j}) = \frac{\alpha_{i,j}/\alpha_{i,0}}{\alpha_{i,0}+1} - \frac{(\alpha_{i,j}/\alpha_{i,0})^2}{\alpha_{i,0}+1}$$

$$Cov(\Delta_{i,j}, \Delta_{i,k}) = -\frac{(\alpha_{i,j}/\alpha_{i,0})(\alpha_{i,k}/\alpha_{i,0})}{\alpha_{i,0}+1}$$

Thus,

$$E((\Delta_i - E(\Delta_i))(\Delta_i - E(\Delta_i))') = \frac{diag(\alpha_i/\alpha_{i,0})}{\alpha_{i,0}+1} - \frac{(\alpha_i/\alpha_{i,0})(\alpha_i/\alpha_{i,0})'}{\alpha_{i,0}+1}$$

where $\alpha_i = (\alpha_{i,1}, \alpha_{i,2}, ..., \alpha_{i,n})$.

3.1 Explicit Moment Conditions

We begin with a more explicit set of conditions to replace (11) to (13). To streamline presentation, from now on we skip the argument $\mathbf{p}$ in $\nabla Z(\mathbf{p})$ and $\nabla^2 Z(\mathbf{p})$ and simply use ∇Z and $\nabla^2 Z$. We have the following:

Theorem 1 If δ satisfies

$$E[\delta] = \mathbf{p} \tag{18a}$$

$$\gamma E[(\delta - \mathbf{p})'(\delta - \mathbf{p})] = I - \frac{\mathbb{1}\mathbb{1}'}{n} \tag{18b}$$

$$E[(\delta - \mathbf{p})_i(\delta - \mathbf{p})_j(\delta - \mathbf{p})_k] = \mu \ \forall\ (i,j,k) \in [n] \times [n] \times [n] \tag{18c}$$

for some $\mu, \gamma \in \mathbb{R}$, assume Z is trice differentiable, and δ has bounded fourth moments, and we let our gradient estimator be

$$\hat{\psi}(\mathbf{p}) = \frac{\gamma}{c}\hat{Z}((1-c)\mathbf{p} + c\delta)(\delta - \mathbf{p})$$

Then

$$E[\hat{\psi}(\mathbf{p})] = \nabla Z + (\frac{c\mu\gamma}{2}\mathbb{1}'\nabla^2 Z\mathbb{1} - \frac{\mathbb{1}'\nabla Z}{n})\mathbb{1} + O(c^2)$$

Proof. Using the Taylor expansion of $Z((1-c)\mathbf{p} + c\delta)$ in (9), we get

$$\begin{aligned} E[\hat{\psi}(\mathbf{p})] &= \frac{\gamma}{c}Z(\mathbf{p})E[(\delta - \mathbf{p})] + \gamma E[(\delta - \mathbf{p})(\delta - \mathbf{p})']\nabla Z + \frac{\gamma c}{2}E[(\delta - \mathbf{p})'\nabla^2 Z(\delta - \mathbf{p})(\delta - \mathbf{p})] + O(c^2) \\ &= (I - \frac{\mathbb{1}\mathbb{1}'}{n})\nabla Z + (\frac{c\mu\gamma}{2}\mathbb{1}'\nabla^2 Z\mathbb{1})\mathbb{1} + O(c^2) \\ &= \nabla Z + (\frac{c\mu\gamma}{2}\mathbb{1}'\nabla^2 Z\mathbb{1} - \frac{\mathbb{1}'\nabla Z}{n})\mathbb{1} + O(c^2) \end{aligned}$$

□

Thus, Theorem 1 implies that if we can enforce the third-order constraint in (18) with $\mu = 0$, then we can have an estimator with bias $O(c^2)$ in estimating $\psi(\mathbf{p})$.

3.2 An Eligible Dirichlet Mixture

We will find a specific implementable δ that satisfies the conditions in (18), by using a mixture of $K = n(n-1)/2+1$ of Dirichlet distributions. Without loss of generality, we assume $\mathbf{p} = (p_1, p_2, ..., p_n)$, $p_1 \leq p_2 \leq ... \leq p_n$. Moreover, each Dirichlet only perturbs two variables, and putting all of them together will have enough degree of freedom to control the covariance matrix and the third-order constraints in (18).

Now consider

$$\delta = \sum_{l=1}^{n-1}\sum_{j=l+1}^{n}\theta^{l,j}\Delta^{l,j} + \theta^n\Delta^n \tag{19}$$

where $\Delta^{l,i} \sim Dir(\alpha_{l,i})$ and $\Delta^n \sim Dir(\alpha_n)$, with

$$\begin{aligned} &\theta^1 = \theta^{1,i} = 2p_1/(n-1),\ i = 2,3,...,n \\ &\theta^2 = \theta^{2,i} = (2p_2 - \theta^1)/(n-2),\ i = 3,4,...,n \\ &... \\ &\theta^l = \theta^{l,i} = (2p_l - \sum_{j=1}^{l-1}\theta^j)/(n-l),\ i = l+1,...,n \\ &... \\ &\theta^{n-1} = \theta^{n-1,n} = (2p_{n-1} - \sum_{j=1}^{n-2}\theta^j) \\ &\theta^n = p_n - \sum_{j=1}^{n-1}\theta^j/2 \end{aligned}$$

and

$$\alpha_{l,i}/\alpha^0_{l,i} = (\mathbf{e}_i + \mathbf{e}_l)/2$$
$$\alpha^0_{l,i} = \mathbb{1}'\alpha_{l,i} = C(\theta^{l,i})^2 - 1$$
$$\alpha_n = \mathbf{e}_n$$

where C is large constant such that $C(\theta^{l,i})^2 > 1$, and $\gamma = 4C/n$.

Theorem 2 The distribution δ chosen as (19) with the depicted parameter choices satisfies (18) with $\mu = 0$.

Theorem 2 stipulates that the choice of δ in (19) can be used in SFE and FFE to obtain a bias $O(c^2)$. Its proofs is skipped due to space limit.

The Dirichlet mixture δ for SFE and FFE works for CFE, assuming the "backward" perturbed distribution $(1+c)\mathbf{p} - c\delta$ is well-defined. Moreover, in such situations, we can obtain a workable δ that mixes only n Dirichlet distributions instead of order n^2, which is depicted as follows. Let $\delta = \sum_{l=1}^n \theta^l \Delta^l,\ \Delta^l \sim Dir(\alpha_l)$ where

$$\theta^l = \begin{cases} np_1 & (l=1) \\ p_l - p_1 & (l = 2,3,...,n) \end{cases}$$

$$\alpha_l = \begin{cases} \mathbb{1}/n & (l=1) \\ \mathbf{e}_l & (l = 2,3,...,n) \end{cases}$$

and $\gamma = \frac{2}{np_1^2}$. We then have

$$E[\delta] = p_1\mathbb{1} + \sum_{l=2}^n (p_l - p_1)\mathbf{e}_l = \sum_{l=1}^n p_l \mathbf{e}_l = \mathbf{p}$$

and

$$\gamma Cov(\delta) = \gamma \frac{np_1^2}{2}\left(diag(\mathbb{1}) - \frac{\mathbb{1}\mathbb{1}'}{n}\right) = I - \frac{\mathbb{1}\mathbb{1}'}{n}$$

We briefly indicate how one may avoid the possibility of the "backward" perturbation lying outside the probability simplex in the CFE. If we start with an initial solution (i.e., distribution) that has positive mass at each support point. Then, by properly tuning the step size and perturbation size at each iteration, the probability mass at each support point can still remain sufficiently positive, so that moving down a mass by the next perturbation size would still be allowable. These choices can be obtained in such a way that still allows convergence under standard assumptions, via an analysis on the constants needed in the scaling of these sizes.

4 NUMERICAL RESULTS

In this section, we conduct numerical experiments to compare the performance of the three gradient estimators. We also experiment the FWSA and its variant, a mirror descent (MD) SA algorithm (Nemirovski et al. 2009; Ghosh and Lam 2015). We consider a deterministic example: $Z(\mathbf{p}) = \mathbf{p}'A'A\mathbf{p} + B'\mathbf{p} + C$ where $\mathbf{p} \in \mathbb{R}^{10}$. (Here we generate $A \in \mathbb{R}^{10\times 10}$, $B \in \mathbb{R}^{10}$, $C \in \mathbb{R}$ randomly such that each entry of them follows $Uniform(0,1)$ independently). Thus, we can evaluate Z exactly and so $var(\hat{Z}) = \sigma^2 = 0$. Note that, though the example is deterministic, the bias and variance discussed in previous sections still hold due to the randomization in the perturbation direction.

We consider solving the minimization problem in (17) with moment constraints

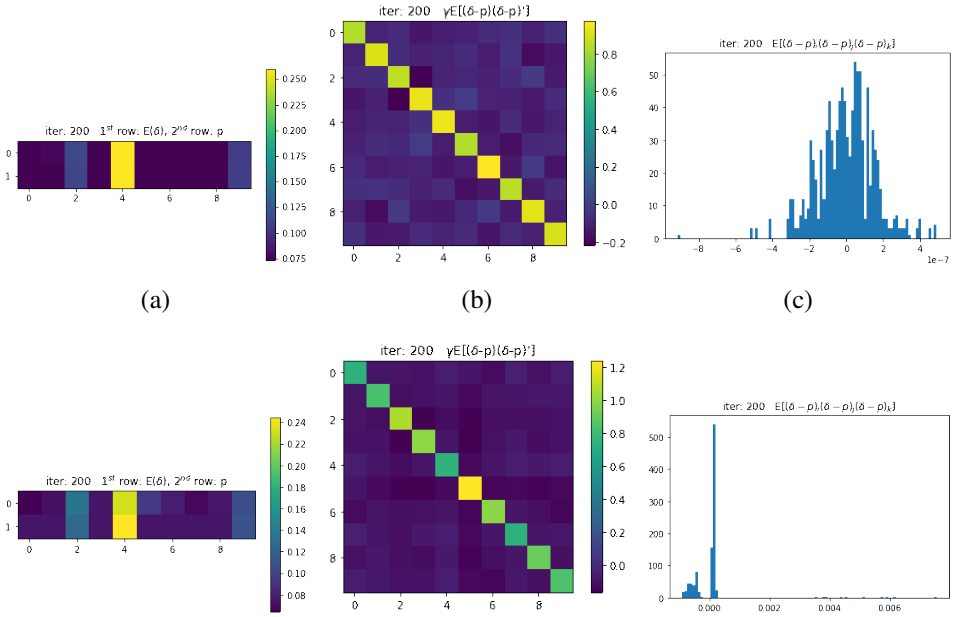
$$\sum_i p_i x_i \in [0.3, 0.7],\ \sum_i p_i x_i^2 \in [0.2815, 0.4222],\ x_i = \frac{i-1}{9},\ i = 1,...,10$$

We choose these bounds because setting $p_1 = ... = p_1 = 1/10$, we have $\sum_i p_i x_i = 0.5, \sum_i p_i x_i^2 \approx 0.3519$, so this set of moment constraints describes a set that is close to the uniform distribution.

The experiment has 2 parts: first, we show that the proposed Dirichlet mixture distribution indeed satisfies the constraints in (18) ((c) for SFE, FFE) and we compare the mean squared error (MSE) of the three gradient estimators under different R and c. Second, we test the proposed FWSA and MDSA algorithm for the convex quadratic function Z.

4.1 Performances of Gradient Estimators

First, we test the distributions for the proposed Dirichlet mixture distributions for Estimator I (SFE), II (FFE) and III (CFE) (For I and II we use the same distribution of δ, so we only show the plot for II, and for III we use the simpler version that only uses n Dirichlets in the mixture). To see this, we show $E[\delta]$, $\gamma E[(\delta-\mathbf{p})(\delta-\mathbf{p})']$, $E[(\delta-\mathbf{p})_i(\delta-\mathbf{p})_j(\delta-\mathbf{p})_k]$ as estimated in the 200th iteration during the SPSA scheme minimizing Z. Here, we choose $R = 500$ for each iteration, and the perturbation size at $k = 200$ is $c_k = \frac{0.1}{k} = 5 \times 10^{-4}$.



(a) (b) (c)

(d) (e) (f)

Figure 1: Distribution of δ. (abc): Estimator I, II; (def): Estimator III.

From Figure 1 we see that for all three estimators, $E[\delta] = \mathbf{p}$ and $\gamma E[(\delta-\mathbf{p})(\delta-\mathbf{p})'] = I - \mathbb{1}\mathbb{1}'/n$ are satisfied. The upper and lower rows in (a) and (d) are almost the same, and the diagonals of (b) and (e) are approximately 0.9, while the off-diagonals are approximately -0.1. By design, however, only Estimator I and II satisfy $E[(\delta-\mathbf{p})_i(\delta-\mathbf{p})_j(\delta-\mathbf{p})_k] = 0$. Indeed, from the histogram over all combination of i, j, k, (c) shows that $(\delta-\mathbf{p})_i(\delta-\mathbf{p})_j(\delta-\mathbf{p})_k \sim 10^{-7}$ are close to 0, while in (f) the values are much larger, and there are outliers. Thus, in Estimator I and II, we need to generate more Dirichlet for each δ, but then the generated δ appears well behaved such that it "cancels out" the $O(c^2)$ terms in the Taylor expansion.

Second, we use Estimators I, II, and III to estimate $\nabla Z(\mathbf{p}_0)$ where $\mathbf{p}_0$ is randomly generated in the probability simplex. We vary c (the perturbation size) and R (the number of δ used in each estimation). We run 100 experiments for each set of parameters and use MSE to measure the performance.

Notice that we know the true gradient at $\mathbf{p}_0$: $\nabla Z(\mathbf{p}_0) = 2A'A\mathbf{p}_0 + B$. And from analysis we know that the bias for each estimator is

$$bias \approx -\frac{\mathbb{1}'\nabla Z(\mathbf{p}_0)}{n}\mathbb{1} = -\frac{\mathbb{1}'(2A'A\mathbf{p}_0 + B)}{n}\mathbb{1}$$

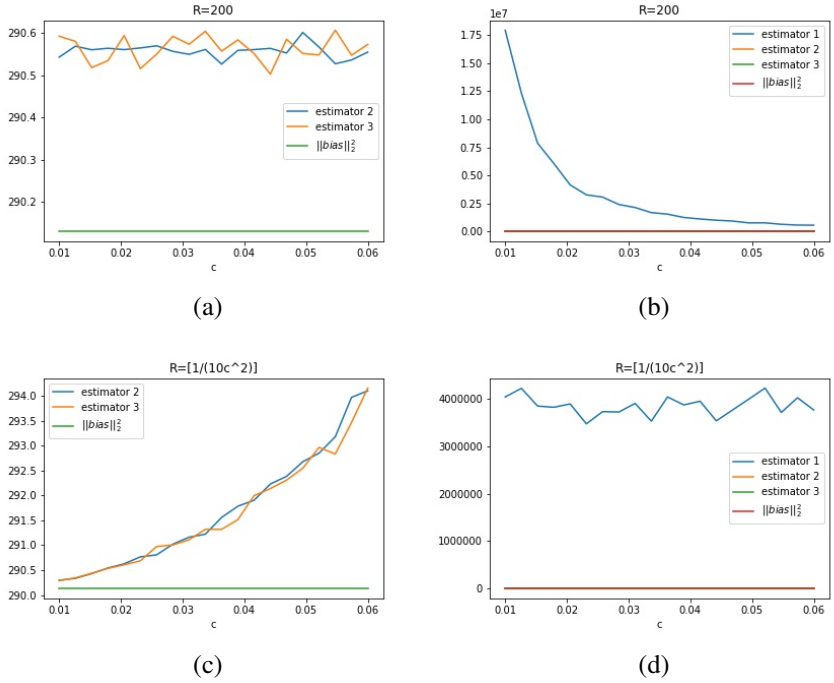


Figure 2: MSE against c.

Figure 2 tells us several things:

1. From Figures (b) and (d), we see that the MSE of Estimator I is much larger than Estimator II and III. In addition, after choosing $R = O(c^{-2})$ as a function of c rather than a constant, the MSE stays at the same level, indicating that its MSE is of order $R^{-1}c^{-2}$.

2. The difference between the MSE and the squared bias can be seen as an estimate of the variance of the estimator. So from figure (a) we see that simply changing c has no effect on variance reduction; while figure (c) shows that to reduce the variance, we can decrease c and increase R together.

3. Based on Figure 2, when choosing which estimator to use in the optimization problem, we should never choose Estimator I due to its large variance. Estimators II and III are comparable in terms of the MSE, but II requires more Dirichlets in the distribution for δ, and III puts constraints on the perturbation size c such that the perturbed $\mathbf{p}$ is still in the simplex. So which one to use might depend on the actual function Z we are interested in. Generally speaking, II is preferable as the overhead in using more Dirichlets is relatively negligible, and its being free of the constraints on c makes it more flexible against III.

4.2 Performances of FWSA and MDSA

In this section, we use FWSA and MDSA to minimize Z. For FWSA, we choose the update stepsize at iteration k to be $\gamma_k = 0.05/k$, the perturbation stepsize $c_k = 0.1/k^{1/4}$. For MDSA, we choose $\gamma_k = 1/k^{1.5}$ in the proximal operator, and the same perturbation size c_k as FWSA. (The MDSA scheme with estimator I gives "nan" during optimization, so we only show the result for FW+Estimator I,II,III and MD+Estimator II,III.) We choose $R = 500$ in each iteration.

We evaluate the performance based on the Frank-Wolfe (FW) gap (for FWSA)

$$g(\mathbf{p}) = -\min_{\mathbf{q}\in\mathscr{U}} \psi(\mathbf{p})'(\mathbf{q}-\mathbf{p})$$

and on the function value Z.

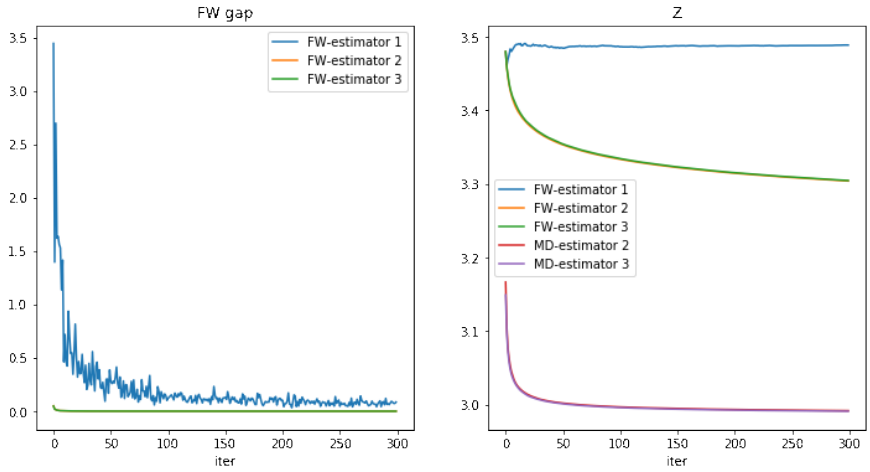


Figure 3: FWSA vs. MDSA.

From Figure 3 we see that all 4 experiments using Estimator II/III are converging, but Estimator I either blows up the optimization (as in MDSA) or does not decrease the function at all (as in FWSA). This might be explained by the poor gradient estimation it provides. However, it is interesting that MDSA converges much faster than FWSA. After 300 iterations, FWSA is still well above MDSA (and is still decreasing). This might be because we did not choose the best set of parameters for the optimization, or might be due to the superiority of MDSA itself. In future work, we plan to test on various functions with different sets of parameters for more extensive investigations.

5 CONCLUSION

In this paper, we investigated gradient estimators using only noisy function evaluations when the input is constrained to be a probability distribution, a problem motivated from running gradient descent for distributionally robust simulation analysis. Traditional random perturbation approaches encounter difficulties in satisfying the conditions needed to cancel out higher-order bias due to the probability simplex constraint. We proposed a new set of conditions for distributionally constrained random perturbation, and studied three types of gradient estimators, SFE, FFE and CFE, by properly mixing the input distribution with a perturbation vector. We showed that while CFE, like in the classical setting, is the easiest to cancel out higher-order bias, the backward perturbation involved may not satisfy the simplex constraint. SFE and FFE can cancel out higher-order bias with more sophisticated perturbation designs, and FFE tends to have a lower variance than SFE due to first-order term canceling. Based on these, we suggested FFE as the recommended choice. We provided two explicit implementations of these estimators using carefully tuned Dirichlet mixtures. Experimental results supported our theoretical findings, and also showed that the proposed estimators are effective for use in FWSA and MDSA. In particular, FFE or CFE together with MDSA appeared to have the best performance in terms of convergence speed.

ACKNOWLEDGMENTS

We gratefully acknowledge support from the National Science Foundation under grants CAREER CMMI-1834710 and IIS-1849280.

REFERENCES

Asmussen, S., and P. W. Glynn. 2007. *Stochastic simulation: algorithms and analysis*, Volume 57. Springer Science & Business Media.

Barton, R. R., B. L. Nelson, and W. Xie. 2014. "Quantifying input uncertainty via simulation confidence intervals". *INFORMS journal on computing* 26(1):74–87.

Barton, R. R., and L. W. Schruben. 2001. "Resampling methods for input modeling". In *Proceeding of the 2001 Winter Simulation Conference (Cat. No. 01CH37304)*, Volume 1, 372–378. IEEE.

Cheng, R. C., and W. Holland. 2004. "Calculation of confidence intervals for simulation output". *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 14(4):344–362.

Chick, S. E. 2001. "Input distribution selection for simulation experiments: accounting for input uncertainty". *Operations Research* 49(5):744–758.

Flaxman, A. D., A. T. Kalai, and H. B. McMahan. 2004. "Online convex optimization in the bandit setting: gradient descent without a gradient". *arXiv preprint cs/0408007*.

Fu, M. C. 2006. "Gradient estimation". *Handbooks in operations research and management science* 13:575–616.

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

Ghosh, S., and H. Lam. 2015. "Mirror descent stochastic approximation for computing worst-case stochastic input models". In *2015 Winter Simulation Conference (WSC)*, 425–436. IEEE.

Ghosh, S., and H. Lam. 2019. "Robust analysis in stochastic simulation: Computation and performance guarantees". *Operations Research* 67(1):232–249.

Glasserman, P. 2013. *Monte Carlo methods in financial engineering*, Volume 53. Springer Science & Business Media.

Glasserman, P., and X. Xu. 2014. "Robust risk measurement and model risk". *Quantitative Finance* 14(1):29–58.

Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. 2011. *Robust statistics: the approach based on influence functions*, Volume 196. John Wiley & Sons.

Kushner, H., and G. G. Yin. 2003. *Stochastic approximation and recursive algorithms and applications*, Volume 35. Springer Science & Business Media.

Lam, H. 2016. "Robust sensitivity analysis for stochastic systems". *Mathematics of Operations Research* 41(4):1248–1275.

Lam, H. 2018. "Sensitivity to serial dependency of input processes: A robust approach". *Management Science* 64(3):1311–1327.

Lam, H., and H. Qian. 2016. "The empirical likelihood approach to simulation input uncertainty". In *2016 Winter Simulation Conference (WSC)*, 791–802. IEEE.

L'Ecuyer, P. 1991. "An overview of derivative estimation". In *Winter Simulation Conference*, 207–217.

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.

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.

Zouaoui, F., and J. R. Wilson. 2003. "Accounting for parameter uncertainty in simulation input modeling". *Iie Transactions* 35(9):781–792.

AUTHOR BIOGRAPHIES

HENRY LAM is an Associate Professor in the Department of Industrial Engineering and Operations Research at Columbia University. He received his Ph.D. degree in statistics from Harvard University in 2011, and was on the faculty of Boston University and the University of Michigan before joining Columbia in 2017. His research focuses on Monte Carlo simulation, uncertainty quantification, risk analysis, and stochastic and robust optimization. His email address is henry.lam@columbia.edu.

JUNHUI ZHANG is an undergraduate junior student at Columbia University, majoring in Applied Mathematics. Her email address is jz2903@columbia.edu.