Proceedings of the 2022 Winter Simulation Conference
B. Feng, G. Pedrielli, Y. Peng, S. Shashaani, E. Song, C.G. Corlu, L.H. Lee, E.P. Chew, T. Roeder, and P. Lendermann, eds.

IMPORTANCE SAMPLING FOR RARE-EVENT GRADIENT ESTIMATION

Yuanlu Bai
Shengyi He
Henry Lam

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

Guangxin Jiang

School of Management
Harbin Institute of Technology
No. 92 Xidazhi Street
Harbin, Heilongjiang 150001, CHINA

Michael C. Fu

Robert H. Smith School of Business and
Institute for Systems Research
University of Maryland
College Park, MD 20742, USA

ABSTRACT

Importance sampling (IS) is a powerful tool for rare-event estimation. However, in many settings, we need to estimate not only the performance expectation but also its gradient. In this paper, we build a bridge from the IS for rare-event estimation to gradient estimation. We establish that, for a class of problems, an efficient IS sampler for estimating the probability of the underlying rare event is also efficient for estimating gradients of expectations over the same rare-event set. We show that both the infinitesimal perturbation analysis and the likelihood ratio estimators can be studied under the proposed framework. We use two numerical examples to validate our findings.

1 INTRODUCTION

Gradient estimation for the performance of stochastic systems is an important topic in simulation analysis, which naturally appears in many problems of interest. A prime example is to solve stochastic optimization problems, such as in simulation-based optimization or empirical risk minimization in machine learning. In these contexts, gradient estimation is used in each iteration of the gradient descent or stochastic approximation algorithm. In addition to optimization, gradient estimation arises in model sensitivity analysis, hedging financial derivatives, among other applications.

Many gradient estimation methods have been proposed (see Fu 2006 for a thorough introduction). If we do not have any information about the system, i.e., the model is "black-box" or that we are in a "zeroth-order" setting, finite-difference approximation is commonly used to estimate gradients. This approach perturbs the value of each component of the parameter of interest (see Glynn 1989; Fox and Glynn 1989; Section 5.2.1 in Fu 2006). If we know some information about the simulation model structure or the input probability distribution, then we can devise unbiased or "first-order" estimators. In particular, infinitesimal perturbation analysis (IPA) or pathwise differentiation, and the likelihood ratio (LR) or the score function

(SF) method, and their variants, have been commonly used. These methods operate by interchanging the derivative and expectation operations. IPA, introduced by Ho and Cao (1983), takes derivative of the sample performance directly, and requires the continuity of the sample performance (Glasserman 1988). To handle discontinuities, smoothed perturbation analysis is proposed to smooth the sample performance by taking a conditional expectation (see Fu and Hu 1997). The LR or SF method, proposed by Reiman and Weiss (1989), Rubinstein (1989) and Glynn (1990), takes derivatives of parameters inside the input distribution and requires the continuity therein (L'Ecuyer 1990). While the LR estimator does not require smoothness of the sample performance, it typically have larger variances than those of IPA. Improvements and complements of the IPA and the LR method include kernel methods (Liu and Hong 2011), weak derivatives (Pflug 1989; Heidergott et al. 2010), and the generalized likelihood ratio method (Wang et al. 2012; Peng et al. 2018).

In this paper, we study gradient estimation where the performance is evaluated on a rare-event region. This problem arises in optimization or sensitivity analysis over tail-related quantities, which appears when training risk-averse predictive models or conducting performance analyses on extremal risk measures (He et al. 2022). In these settings, estimation based on crude Monte Carlo (MC), even just for expectation evaluations, would be inefficient, because a prohibitively large sample size is required to ensure the rare event is observed sufficiently adequately. Statistically, this inefficiency manifests as a huge estimation variance. To address this challenge, variance reduction techniques such as importance sampling (IS), are known to be powerful in enhancing efficiency. The idea of IS is to draw samples from an alternative distribution that hits the rare event more frequently and adjusts the estimator with a likelihood ratio to retain unbiasedness. When the alternative distribution is properly chosen, IS can significantly reduce the variance of the estimation. Selecting a good alternative distribution has been the key of IS and has been extensively studied (see, e.g., the surveys Bucklew 2004; Juneja and Shahabuddin 2006; Blanchet and Lam 2012).

Despite a vast literature on the application of IS to rare-event estimation problems, there are few studies on using IS for gradient estimation problems that involve rare events which, as mentioned above, nonetheless arises in risk-averse model training and performance analysis. In this paper, we study rare-event gradient estimation by looking at a class of problems that generically arise in this context. More concretely, we consider estimating the expectation of an objective function that is piecewise polynomial, evaluated on a rare event that is defined as the excursion of a piecewise linear function applied on a random object. Under suitable conditions, this objective function can result from both the IPA and the LR or SF methods in estimating the gradients of rare-event probabilities or related risk quantities. We show that in the specific case of Gaussian inputs, the IS based on Gaussian mixture designed for the estimation of the underlying rare-event probability is also efficient for the estimation of the more general objective function. This suggests that once we obtain an efficient sampler for the rare-event probability, the same sampler could be used for a wide class of gradient estimation problems on the same rare-event set. Our result thus facilitates the design of rare-event gradient estimators by identifying a readily obtainable efficient IS. Although our results only focus on Gaussian inputs, similar theories are conjectured to hold for more general light-tailed distributions which we relegate to a full journal version of this work.

We briefly review other studies on variance reduction techniques for gradient estimation. The most relevant work we have seen is Nakayama (1995), which studies the gradients on the performances of highly reliable Markovian systems. It is shown that when applying a type of IS called "balanced failure biasing", all gradients can be estimated as accurately as the performance measure itself. Our study can be viewed as a generalization of this observation. Besides Nakayama (1995), control variates have been applied in stochastic gradient descent algorithms to reduce the variance of the gradient estimate and improve the convergence rates of the algorithms (see Wang et al. 2013 and Driggs et al. 2022). For empirical risk minimization problems, stochastic variance reduced gradient (SVRG) was proposed, which progressively reduces the variance using the averaged gradient at an estimated optimal solution. To save computational cost, the estimated optimal solution is updated only periodically (Johnson and Zhang 2013; Xiao and Zhang

2014). Another work on variance reduction in empirical risk minimization is Zhao and Zhang (2015), which uses IS to select the data when running stochastic gradient descent. However, as mentioned before, there are few works on IS to reduce the variance of the gradient estimator with rare events, even though IS has been extensively studied in estimating a rare-event probability or its related expectation.

The remainder of this paper is organized as follows: Section 2 introduces our problem class and the proposed IS for gradient estimation, and shows the efficiency of the proposed IS. In Section 3, we use two numerical examples to show the efficiency of the mixture IS in the estimation of both rare-event expectation and gradient. Finally, Section 4 provides concluding remarks.

2 IMPORTANCE SAMPLING FOR GRADIENT ESTIMATION

As we will show momentarily, a class of gradients derived using either IPA or LR can be formulated as

$$\mathbb{E}\left[f(X)1\{g(X)\geq\gamma\}\right], \tag{1}$$

where $\{g(X)\geq\gamma\}$ represents a rare event. We impose the following assumption regarding the underlying distribution, the function f and the structure of the rare event.

Assumption 1 $X\sim N(\mu,\Sigma)$ for some $\mu\in\mathbb{R}^d$ and $\Sigma\in\mathbb{R}^{d\times d}$. $f(\cdot)$ is a piecewise polynomial and $g(\cdot)$ is a piecewise linear function such that we can split $\{g(x)\geq\gamma\}$ into n mutually exclusive polyhedrons $\{A^{(i)}x+b^{(i)}\geq t^{(i)}\}(i=1,\cdots,n)$ and $f(x)=f_i(x)$ over polyhedron $\{A^{(i)}x+b^{(i)}\geq t^{(i)}\}$, where $f^{(i)}(\cdot)$ is a polynomial function. Moreover, for any $i=1,2,\ldots,n$, $\{A^{(i)}x+b^{(i)}\geq t^{(i)}\}$ is full-dimensional, i.e., $P(\{A^{(i)}X+b^{(i)}\geq t^{(i)}\})>0$ where $X\sim N(\mu,\Sigma)$.

We show that a class of IPA and LR can be formulated as (1) such that Assumption 1 holds.

Example 1 (IPA) Suppose we are interested in the gradient of $\mathbb{E}\left[f(X,\theta)1\{g(X)\geq\gamma\}\right]$ w.r.t. θ, where $X\sim N(\mu,\Sigma)$ and $f(x,\theta)$ is a piecewise polynomial function of x whose coefficients are determined by θ. Then we have that

$$\frac{\partial}{\partial\theta}\mathbb{E}\left[f(X,\theta)1\{g(X)\geq\gamma\}\right]=\mathbb{E}\left[\frac{\partial}{\partial\theta}f(X,\theta)1\{g(X)\geq\gamma\}\right],$$

which has the form of (1).

Example 2 (LR) Suppose we are interested in the gradient $\frac{\partial}{\partial\theta}\mathbb{E}_\theta\left[f(X)1\{g(X)\geq\gamma\}\right]$ where $X\sim N(\mu(\theta),\Sigma)$. Using LR and noting that

$$\frac{\partial}{\partial\theta}\log p_\theta(X)=\frac{\partial}{\partial\theta}\mu(\theta)^\top\Sigma^{-1}(X-\mu(\theta)),$$

where $p_\theta(X)$ is the density of $N(\mu(\theta),\Sigma)$, we have

$$\frac{\partial}{\partial\theta}\mathbb{E}_\theta\left[f(X)1\{g(X)\geq\gamma\}\right]=\mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\mu(\theta)^\top\Sigma^{-1}(X-\mu(\theta))\right)f(X)1\{g(X)\geq\gamma\}\right].$$

If f is a piecewise polynomial, then the function $\left(\frac{\partial}{\partial\theta}\mu(\theta)^\top\Sigma^{-1}(X-\mu(\theta))\right)f(X)$ is also a piecewise polynomial. Therefore, the LR estimator can also be written as the form of (1).

For the estimation of the probability $P(g(X)\geq\gamma)$, an efficient sampler is given in Bai et al. (2022), which is described as follows. For each $i=1,2,\ldots,n$, let x_i^* denote the minimizer of $(x-\mu)^\top\Sigma^{-1}(x-\mu)$ over the polyhedron $\{A^{(i)}X+b^{(i)}\geq t^{(i)}\}$. The proposed sampler is a Gaussian mixture whose density is

$$p(x)=\frac{1}{n}\sum_{i=1}^{n}p_i(x) \tag{2}$$

where $p_i(x)$ is the density of $N(x_i^*, \Sigma)$. The associated likelihood ratio is given by

$$L(x) = \frac{ne^{-\frac{1}{2}(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)}}{e^{-\frac{1}{2}(x-x_1^*)^{\mathrm{T}}\Sigma^{-1}(x-x_1^*)} + \cdots + e^{-\frac{1}{2}(x-x_n^*)^{\mathrm{T}}\Sigma^{-1}(x-x_n^*)}}.$$

We propose to use the same sampler for the problem (1). Therefore, the proposed estimator for (1) is given by

$$f(\tilde{X})L(\tilde{X})1\{g(\tilde{X}) \geq \gamma\},$$

where $\tilde{X} \sim p(\cdot)$. We will show that the proposed estimator is efficient. Typically, for rare-event estimation problems, the efficiency is defined relative to the mean of the estimator. However, since we are looking at gradient estimation problems, usually we do not have $f > 0$ in (1), so it might happen that the mean is extremely small due to the cancellation between the positive parts of f and the negative parts of f. For this reason, we use the rare-event probability as the benchmark. This benchmark is reasonable, although its usefulness would depend on the nature of the downstream tasks in consideration. More precisely, we have the following theorem.

Theorem 1 Suppose that Assumption 1 holds. Let $\tilde{X}$ be a sample drawn from $p(\cdot)$ as defined in (2) and $X \sim N(\mu, \Sigma)$. Then we have that

$$\frac{\mathbb{E}\left[f(\tilde{X})^2 L^2(\tilde{X})1\{g(\tilde{X}) \geq \gamma\}\right]}{(P(g(X) \geq \gamma))^2} \tag{3}$$

does not grow exponentially as $\gamma \to \infty$.

Proof. Without loss of generality, throughout the proof we suppose that $\mu = 0$. Under Assumption 1, the expectation in (1) can be written as

$$\sum_{i=1}^{n} \mathbb{E}\left[f_i(X)1\{A^{(i)}X + b^{(i)} \geq t^{(i)}\}\right].$$

For each $i = 1, 2, \ldots, n$, let $N(x_i^*) = \{\|x - x_i^*\|_2 \leq \varepsilon\}$ denote a neighborhood of x_i^*. Since $\{A^{(i)}x + b^{(i)} \geq t^{(i)}\}$ is full-dimensional, by introducing redundant constraints, we may find a full-rank $B^{(i)}$ and a vector $s^{(i)}$ such that $\{A^{(i)}x + b^{(i)} \geq t^{(i)}\} \subset \{B^{(i)}x \geq s^{(i)}\}$ and when ε is sufficiently small,

$$N(x_i^*) \cap \{A^{(i)}x + b^{(i)} \geq t^{(i)}\} = N(x_i^*) \cap \{B^{(i)}x \geq s^{(i)}\}.$$

See Lemma 7.2 and Lemma 7.1 of Bai et al. (2022) for more details on the construction of $B^{(i)}$ and $s^{(i)}$. From Lemma 7.2 of Bai et al. (2022), we have that

$$P(\{A^{(i)}X + b^{(i)} \geq t^{(i)}\}) \overset{poly}{\sim} e^{-\frac{1}{2}x_i^{*\top}\Sigma^{-1}x_i^*} \tag{4}$$

where $f_1(\gamma) \overset{poly}{\sim} f_2(\gamma)$ if $f_1(\gamma)/f_2(\gamma)$ changes at most polynomially in γ.

The second moment of the estimator can be written as

$$\mathbb{E}[(f(\tilde{X}))^2 1\{g(\tilde{X}) \geq \gamma\}L^2(\tilde{X})] = \mathbb{E}[(f(X))^2 1\{(g(X) \geq \gamma\}L(X)] = \sum_{i=1}^{n} \mathbb{E}[(f_i(X))^2 L(X)1\{A^{(i)}X + b^{(i)} \geq t^{(i)}\}].$$

For each i, we are going to give an upper bound of $\mathbb{E}[(f_i(X))^2 L(X)1\{A^{(i)}X + b^{(i)\}} \geq t^{(i)}]$. We know that

$$L(x) \leq \frac{ne^{-\frac{1}{2}x^{\mathrm{T}}\Sigma^{-1}x}}{e^{-\frac{1}{2}(x-x_i^*)^{\mathrm{T}}\Sigma^{-1}(x-x_i^*)}} = ne^{\frac{1}{2}x_i^{*\top}\Sigma^{-1}x_i^* - x^{\mathrm{T}}\Sigma^{-1}x_i^*}.$$

We omit the index i in $A^{(i)}, B^{(i)}, b^{(i)}, s^{(i)}, x_i^*$ from now on for convenience. Let $\phi_\Sigma(x)$ denote the density of $N(0,\Sigma)$. Using the above inequality, we have that

$$\begin{aligned}
&\mathbb{E}[(f_i(X))^2 L(X) 1\{AX+b \geq t\}] \leq \mathbb{E}[(f_i(X))^2 L(X) 1\{BX \geq s\}] = \int_{\{Bx \geq s\}} (f_i(x))^2 L(x) \phi_\Sigma(x) \mathrm{d}x \\
\leq & \int_{\{Bx \geq s\}} n e^{\frac{1}{2} x^{*\top} \Sigma^{-1} x^* - x^{\mathrm{T}} \Sigma^{-1} x^*} (f_i(x))^2 \phi_\Sigma(x) \mathrm{d}x \\
= & n \left(\frac{1}{2\pi|\Sigma|}\right)^{\frac{d}{2}} \int_{\{Bx \geq s\}} (f_i(x))^2 e^{-\frac{1}{2}(x-x^*)^{\mathrm{T}} \Sigma^{-1} (x-x^*) - (x-x^*)^{\mathrm{T}} \Sigma^{-1} x^* - x^{\mathrm{T}} \Sigma^{-1} x^*} \mathrm{d}x \\
= & n \left(\frac{1}{2\pi|\Sigma|}\right)^{\frac{d}{2}} e^{-x^{*\top} \Sigma^{-1} x^*} \int_{\{Bx \geq s\}} (f_i(x))^2 e^{-\frac{1}{2}(x-x^*)^{\mathrm{T}} \Sigma^{-1} (x-x^*) - 2(x-x^*)^{\mathrm{T}} \Sigma^{-1} x^*} \mathrm{d}x \\
\leq & n \left(\frac{1}{2\pi|\Sigma|}\right)^{\frac{d}{2}} e^{-x^{*\top} \Sigma^{-1} x^*} \int_{\{Bx \geq s\}} (f_i(x))^2 e^{-\frac{1}{2}(x-x^*)^{\mathrm{T}} \Sigma^{-1} (x-x^*)} \mathrm{d}x \\
\leq & n \frac{1}{|B|} \left(\frac{1}{2\pi|\Sigma|}\right)^{\frac{d}{2}} e^{-x^{*\top} \Sigma^{-1} x^*} \int_{\{y \geq s - Bx^*\}} (f_i(x^* + B^{-1} y))^2 e^{-\frac{1}{2} \lambda_{min} y^{\mathrm{T}} y} \mathrm{d}y.
\end{aligned}$$

Here we used the substitution $x = x^* + B^{-1} y$ in the last step and λ_{min} is the minimum eigenvalue of $(B^{-1})^\top \Sigma^{-1} B^{-1}$. Since f_i is a polynomial, $\int_{\{y \geq s - Bx^*\}} (f_i(x^* + B^{-1} y))^2 e^{-\frac{1}{2} \lambda_{min} y^{\mathrm{T}} y} \mathrm{d}y$ can be bounded from above by a polynomial of absolute values of components of x^*. Therefore, combining the above inequality with (4), we get the desired result. □

In Bai et al. (2022), the same sampler is shown to be relatively efficient for the estimation of the probability of the rare event $\{g(X) \geq \gamma\}$. Combining this with Theorem 1, we observe that in this case, an efficient sampler for the underlying rare-event probability is also efficient for the estimation of gradients of a wide class of expectations defined on the same rare-event set that satisfies Assumption 1. A generalization of this observation to settings beyond Assumption 1 (e.g., models with general light-tailed underlying distributions and general performance functions) is left as future work. Another direction of our future work is to give a tighter bound for the growth of (3) as γ grows in terms of a polynomial whose degree and coefficients are determined by model parameters.

In many applications, we want the gradient of expectations of positive functions. An example is expected loss which is typically positive over the entire region in consideration. In this case, with Theorem 1, we can also show relative efficiency using the expectation before taking derivative as the benchmark.

Corollary 2 (IPA) In Example 1, if $f(x,\theta) > 0$ for all $x \in \mathbb{R}^d$ and Assumption 1 holds when the problem is formulated as (1), then we have

$$\frac{\mathbb{E}\left[\left(\frac{\partial}{\partial \theta} f(\tilde{X}, \theta)\right)^2 L^2(\tilde{X}) 1\{g(\tilde{X}) \geq \gamma\}\right]}{\left(\mathbb{E}\left[f(\tilde{X}, \theta) L(\tilde{X}) 1\{g(\tilde{X}) \geq \gamma\}\right]\right)^2}$$

does not grow exponentially as $\gamma \to \infty$.

Corollary 3 (LR) In Example 2, suppose that $f(x) > 0$ for all $x \in \mathbb{R}^d$ and Assumption 1 holds when the problem is formulated as (1). Let $\tilde{X}$ be a sample drawn from $p(\cdot)$ as defined in (2) with $\mu = \mu(\theta)$, then we have

$$\frac{\mathbb{E}\left[\left(\frac{\partial}{\partial \theta} \mu(\theta)^\top \Sigma^{-1} (\tilde{X} - \mu(\theta)) f(\tilde{X})\right)^2 L^2(\tilde{X}) 1\{g(\tilde{X}) \geq \gamma\}\right]}{\left(\mathbb{E}\left[f(\tilde{X}) L(\tilde{X}) 1\{g(\tilde{X}) \geq \gamma\}\right]\right)^2}$$

does not grow exponentially as $\gamma \to \infty$.

3 NUMERICAL EXPERIMENTS

In this section, we consider two toy examples and a finance example (pricing a basket option and its delta) to demonstrate the effectiveness of the IS approach in both the performance function and gradient estimation.

3.1 Toy Examples

In the first toy example, we verify the effectiveness of IS for IPA. Suppose that $X = (X_1, X_2)^\top \sim N(0, I_2)$ where I_2 is the 2×2 identity matrix. The performance function is

$$f(X, \theta) = \theta^2 X_1^2 + X_2^2 + 0.01,$$

and the constraint function is

$$g(X) = \max\{2X_1 + X_2, X_2 + 2X_2\}. \tag{5}$$

We would like to estimate $\mathbb{E}[f(X,\theta)1\{g(X) \geq \gamma\}]$. Specifically, we set $\gamma = 7$ and $\theta = 2$, and use the IPA estimator given by $2\theta X_1^2 1\{g(X) \geq \gamma\}$.

According to (2) and (5), we obtain $x_1^* = (1.4, 2.8)^\top$ and $x_2^* = (2.8, 1.4)^\top$. Therefore, the IS distribution is a mixture of $N(x_1^*, I_2)$ with probability 50% and $N(x_2^*, I_2)$ with probability 50%. For sample size from 1000 to 500000, Table 1 shows that the proposed mixture IS works well both for the performance and

Table 1: Means, variances, and variance reduction ratios of performance and gradient estimation (IPA) for MC and IS in the first toy example (true performance and gradient are 4.72E-02 and 3.78E-02, respectively).

	Performance					Gradient				
Size	IS		MC			IS		MC		
	mean	variance	mean	variance	ratio	mean	variance	mean	variance	ratio
1000	4.70E-02	8.31E-06	4.89E-02	1.92E-03	231.3	3.77E-02	7.04E-06	3.88E-02	1.51E-03	214.1
2000	4.72E-02	3.94E-06	5.03E-02	9.14E-04	231.7	3.78E-02	3.53E-06	4.02E-02	7.36E-04	208.8
5000	4.72E-02	2.01E-06	4.82E-02	3.35E-04	166.4	3.77E-02	1.74E-06	3.89E-02	2.84E-04	163.4
10000	4.71E-02	8.28E-07	4.60E-02	1.82E-04	219.8	3.77E-02	7.40E-07	3.66E-02	1.48E-04	200.0
20000	4.72E-02	4.51E-07	4.73E-02	7.58E-05	168.0	3.77E-02	3.88E-07	3.79E-02	6.35E-05	163.5
50000	4.72E-02	1.44E-07	4.71E-02	3.97E-05	274.7	3.78E-02	1.30E-07	3.79E-02	3.43E-05	263.2
100000	4.72E-02	7.72E-08	4.73E-02	1.83E-05	236.5	3.77E-02	7.14E-08	3.79E-02	1.45E-05	203.5
200000	4.72E-02	5.33E-08	4.69E-02	9.61E-06	180.2	3.78E-02	4.55E-08	3.75E-02	8.11E-06	178.1
500000	4.72E-02	1.61E-08	4.72E-02	3.66E-06	227.8	3.78E-02	1.38E-08	3.78E-02	2.89E-06	209.2

gradient estimation (via IPA). For example, when the sample size is 500000, the variance reduction ratios for performance and gradient estimation are 227.8 and 209.2, respectively. When the sample size is small, e.g., less than 10000, the IS for both performance and gradient estimation has a good accuracy, while classical MC does not.

Furthermore, we change the rare-event parameter γ from 4 to 11 with sample size 100000, and calculate the efficiency ratios of performance estimation and gradient estimation, which are given by

$$\frac{\mathrm{Var}\left(f(X,\theta)1\{g(X) \geq \gamma\}\right)}{\left(\mathbb{E}[f(X,\theta)1\{g(X) \geq \gamma\}]\right)^2} \quad \text{and} \quad \frac{\mathrm{Var}\left(\frac{\partial}{\partial\theta} f(X,\theta)1\{g(X) \geq \gamma\}\right)}{\left(\mathbb{E}[f(X,\theta)1\{g(X) \geq \gamma\}]\right)^2}, \tag{6}$$

respectively, for classical MC, and

$$\frac{\mathrm{Var}\left(f(\tilde{X},\theta)L(\tilde{X})1\{g(\tilde{X}) \geq \gamma\}\right)}{\left(\mathbb{E}[f(\tilde{X},\theta)L(\tilde{X})1\{g(\tilde{X}) \geq \gamma\}]\right)^2} \quad \text{and} \quad \frac{\mathrm{Var}\left(\frac{\partial}{\partial\theta} f(\tilde{X},\theta)L(\tilde{X})1\{g(\tilde{X}) \geq \gamma\}\right)}{\left(\mathbb{E}[f(\tilde{X},\theta)L(\tilde{X})1\{g(\tilde{X}) \geq \gamma\}]\right)^2}, \tag{7}$$

respectively, for IS. The results are shown in Table 2, which indicates that, as γ increases, the efficiency ratios of performance and gradient estimation will blow up for MC, while they grow very slowly for IS.

Table 2: Efficiency ratios in the first toy example (sample size 10^5).

γ		4	5	6	7	8	9	10	11
Objective	IS	2.31E+00	2.53E+00	3.19E+00	3.88E+00	3.72E+00	4.91E+00	5.10E+00	6.32E+00
	MC	2.08E+01	5.56E+01	2.07E+02	8.18E+02	3.63E+03	1.97E+04	1.35E+05	9.72E+05
Gradient	IS	2.18E+00	2.37E+00	2.80E+00	3.54E+00	3.08E+00	4.14E+00	4.26E+00	5.17E+00
	MC	1.79E+01	4.73E+01	1.64E+02	6.46E+02	2.95E+03	1.66E+04	1.08E+05	7.61E+05

In the second toy example, we verify the effectiveness of IS for LR. Suppose that the performance function is

$$f(X) = X_1^2 + X_2^2 + 0.01,$$

where $X = (X_1, X_2) \sim N([\theta, 0]^\top, I_2)$. The constraint function is the same as (5). Specifically, we set $\gamma = 7$ and $\theta = 0$, and the LR estimator is given by $f(X)(X_1 - \theta)1\{g(X) \geq \gamma\}$.

Similar to the first toy example, we vary the sample size from 1000 to 500000 with $\gamma = 7$ and vary the rare-event parameter γ from 4 to 11 with sample size 100000, respectively, to obtain Tables 3 and 4. From these tables, we observe the similar results as in the first toy example: The mixture IS works well for both the performance and gradient estimation (via LR). As the rare-event parameter grows, the efficiency ratios of performance and gradient estimation for IS grow very slowly, while they blow up for classical MC.

Table 3: Means, variances, and variance reduction ratios of performance and gradient estimation (LR) for MC and IS in the second toy example (true objective and gradient is 1.89E-02 and 4.25E-02, respectively).

	Objective					Gradient				
Size	IS		MC			IS		MC		
	mean	variance	mean	variance	ratio	mean	variance	mean	variance	ratio
1000	1.89E-02	1.10E-06	1.76E-02	2.28E-04	207.9	4.25E-02	7.13E-06	3.97E-02	1.33E-03	186.5
2000	1.89E-02	5.07E-07	1.81E-02	9.74E-05	192.2	4.25E-02	3.29E-06	4.16E-02	6.82E-04	207.7
5000	1.89E-02	2.23E-07	1.88E-02	5.22E-05	234.1	4.26E-02	1.44E-06	4.24E-02	3.09E-04	214.5
10000	1.89E-02	1.17E-07	1.84E-02	2.49E-05	213.0	4.25E-02	7.19E-07	4.13E-02	1.72E-04	238.5
20000	1.89E-02	5.29E-08	1.92E-02	1.14E-05	214.6	4.25E-02	3.70E-07	4.35E-02	8.48E-05	229.2
50000	1.89E-02	2.17E-08	1.88E-02	4.46E-06	206.1	4.25E-02	1.39E-07	4.21E-02	2.90E-05	209.3
100000	1.89E-02	1.29E-08	1.90E-02	2.32E-06	180.6	4.25E-02	7.88E-08	4.25E-02	1.60E-05	203.0
200000	1.89E-02	5.43E-09	1.89E-02	1.40E-06	257.7	4.25E-02	3.85E-08	4.26E-02	9.30E-06	241.7
500000	1.89E-02	1.91E-09	1.88E-02	4.31E-07	226.1	4.25E-02	1.24E-08	4.25E-02	3.46E-06	278.7

Table 4: Efficiency ratios in the second toy example (sample size 10^5).

γ		4	5	6	7	8	9	10	11
Objective	IS	1.62E+00	1.94E+00	2.51E+00	2.85E+00	3.87E+00	4.68E+00	4.06E+00	6.29E+00
	MC	1.82E+01	5.13E+01	1.57E+02	7.54E+02	3.17E+03	1.57E+04	1.52E+05	1.29E+06
Gradient	IS	5.12E+00	8.90E+00	1.30E+00	1.86E+01	2.97E+01	4.43E+01	4.16E+01	7.90E+01
	MC	8.17E+01	2.42E+02	7.50E+02	4.98E+03	2.58E+04	1.69E+05	1.67E+06	1.92E+07

3.2 Basket Option

In this example, we estimate the price of a basket option and its *delta*, i.e., the partial derivative of the option price w.r.t. the underlying asset prices. Suppose that there is a basket option with two underlying assets, which are driven by geometric Brownian motions $S_1(t)$ and $S_2(t)$, i.e.,

$$S_i(T) = S_i(0)\exp\left((r - \frac{1}{2}\sigma_i^2)T + \sigma_i W_i(T)\right) = S_i(0)\exp\left((r - \frac{1}{2}\sigma_i^2)T + \sigma_i\sqrt{T}X_i\right), \tag{8}$$

where $S_i(T)$ is the price of underlying asset i at time T, r is the risk-free interest rate, σ_i is the volatility of asset i, and $W_i(t), i = 1,2$ are standard Brownian motions, which are mutually independent, i.e., $(X_1, X_2)^\top \sim N(0, I_2)$. Let $X = (X_1, X_2)^\top$ and $S(T) = (S_1(T), S_2(T))^\top$.

We consider a geometric-average basket option with strike price K, i.e., the payoff function is

$$V(S(T),K) = \left(\left(\prod_{i=1}^{2} S_i(T) \right)^{\frac{1}{2}} - K \right) 1\left\{ \left(\prod_{i=1}^{2} S_i(T) \right)^{\frac{1}{2}} \geq K \right\}. \tag{9}$$

The performance function of interest is $\mathbb{E}[V(S(T),K)]$. When K is large, $\{\left(\prod_{i=1}^{2} S_i(T)\right)^{1/2} \geq K\}$ is a rare-event set, and can be reformulated as

$$\begin{aligned} \left(\prod_{i=1}^{2} S_i(T) \right)^{\frac{1}{2}} &= \left(\prod_{i=1}^{2} S_i(0) \exp\left((r - \frac{1}{2}\sigma_i^2)T + \sigma_i \sqrt{T} X_i \right) \right)^{\frac{1}{2}} \\ &= (S_1(0)S_2(0))^{\frac{1}{2}} \exp\left(\left(r - \frac{\sigma_1^2 + \sigma_2^2}{4} \right) T + \frac{1}{2}\sqrt{T}(\sigma_1 X_1 + \sigma_2 X_2) \right) \geq K. \end{aligned}$$

Therefore, setting

$$g(X) \triangleq \frac{1}{2}\sqrt{T}\sigma^\top X + \frac{1}{2}(\log S_1(0) + \log S_2(0)) + \left(r - \frac{\sigma_1^2 + \sigma_2^2}{4} \right) T,$$

where $\sigma = (\sigma_1, \sigma_2)^\top$, the rare event can be written as $\{g(X) \geq \log K\}$. Note that $g(X)$ is a linear function.

The objective function in (9) is given in terms of X as

$$\left(\prod_{i=1}^{2} S_i(T) \right)^{\frac{1}{2}} - K = (S_1(0)S_2(0))^{\frac{1}{2}} \exp\left(\left(r - \frac{\sigma_1^2 + \sigma_2^2}{4} \right) T + \frac{1}{2}\sqrt{T}(\sigma^\top X) \right) - K. \tag{10}$$

We note that the above is exponential in X and is thus heavy-tailed, which does not satisfy Assumption 1. But when K is not too large, we expect that a Taylor series expansion could approximate the original objective function reasonably well. This approximation is heuristic, but from the numerical examples as we will show, the approximated objective is quite close to the original quantity. More specifically, let $\beta \triangleq (S_1(0)S_2(0))^{1/2} \exp\left(\left(r - (\sigma_1^2 + \sigma_2^2)/4 \right) T \right)$. Then the payoff function in (9) can be approximated by a l-th order Taylor series expansion:

$$\begin{aligned} \left(\prod_{i=1}^{2} S_i(T) \right)^{\frac{1}{2}} - K &= (S_1(0)S_2(0))^{\frac{1}{2}} \exp\left(\left(r - \frac{\sigma_1^2 + \sigma_2^2}{4} \right) T + \frac{1}{2}\sqrt{T}(\sigma_1 X_1 + \sigma_2 X_2) \right) - K \\ &= \beta \exp\left(\frac{1}{2}\sqrt{T}\sigma^\top X \right) - K \\ &\approx \beta - K + \beta \left(\frac{1}{2}\sqrt{T}\sigma^\top X + \frac{1}{2^2 2!} T (\sigma^\top X)^2 + \cdots + \frac{1}{2^l l!} T^{\frac{l}{2}} (\sigma^\top X)^l \right) \\ &\triangleq f(X). \end{aligned} \tag{11}$$

Furthermore, letting $A = \sqrt{T}\sigma/2$, $b = (\log S_1(0) + \log S_2(0))/2 + \left(r - (\sigma_1^2 + \sigma_2^2)/4 \right) T$, and $t = \log K$, then the price of the basket option (without discount factor) can be approximated by

$$\mathbb{E}\left[\left(\left(\prod_{i=1}^{2} S_i(T) \right)^{\frac{1}{2}} - K \right) 1\left\{ \left(\prod_{i=1}^{2} S_i(T) \right)^{\frac{1}{2}} \geq K \right\} \right] \approx \mathbb{E}\left[f(X) 1\{A^\top X + b \geq t\} \right].$$

Note that the RHS above has the form of (1), which satisfies Assumption 1. Therefore, the IS distribution is chosen as $N(x^*, I_2)$, where x^* is the minimizer of $\|x\|_2$ over polyhedron $\{AX + b \geq t\}$ given by

$$x^* = \frac{t-b}{\|A\|^2} A. \tag{12}$$

Using this IS distribution, we would sample $\tilde{X} = (\tilde{X}_1, \tilde{X}_2)^\top \sim N(x^*, I_2)$. Then we can compute $\tilde{S}_i(T)$, which is the counterpart of $S_i(T)$ driven by the normal random variable $\tilde{X}_i$ based on (8). Let $L(\tilde{X})$ be the likelihood ratio. We have that

$$\mathbb{E}[V(S(T), K)] \approx \mathbb{E}[f(X) 1\{g(X) \geq\} \log K] = \mathbb{E}[f(\tilde{X}) 1\left\{g(\tilde{X}) \geq \log K\right\} L(\tilde{X})].$$

3.2.1 Gradient Estimation

For gradient estimation, we use the IPA to derive the delta estimator. We consider both the original objective function given in (9) and the approximated function given in (11). For the original objective function, the IPA estimator of the delta with respect to $S_i(0)$ is

$$\frac{1}{2S_i(0)} \left(\prod_{i=1}^{2} S_i(T) \right)^{\frac{1}{2}} 1\left\{g(X) \geq \log K\right\},$$

and the IPA estimator with IS is

$$\frac{1}{2S_i(0)} \left(\prod_{i=1}^{2} \tilde{S}_i(T) \right)^{\frac{1}{2}} 1\left\{g(\tilde{X}) \geq \log K\right\} L(\tilde{X}).$$

For the approximated function (11), the IPA estimator for classical MC is

$$\frac{1}{2S_i(0)} \left(f(X) + K\right) 1\left\{g(X) \geq \log K\right\},$$

and the IPA estimator with IS is

$$\frac{1}{2S_i(0)} \left(f(\tilde{X}) + K\right) 1\left\{g(\tilde{X}) \geq \log K\right\} L(\tilde{X}).$$

3.2.2 Simulation Results

We investigate the variance reduction effects both for the approximated objective function (11) that satisfies the polynomial setting in the paper and the original exponential objective function (10) that does not satisfy the piece-wise polynomial setting. Let $S_1(0) = 50$, $S_2(0) = 60$, $\sigma_1 = 0.2$, $\sigma_2 = 0.15$, $T = 0.1$, and $K = 60$. For the approximated objective function (11), we set $l = 4$, i.e., we use a fourth-order Taylor series expansion (a fourth-order polynomial) to approximate the true objective function. We vary the total simulation budget from 1000 to 500000, and obtain the following tables.

From Tables 5 and 6, we observe the following: (i) The fourth-order polynomial is quite accurate in approximating the original exponential objective function and also works in the delta estimation. In Table 5, the price means of IS for the fourth-order polynomial approximation are all about 8.90E-03, which is the true price according to the Black-Scholes formula. And the price means of MC for the fourth-order polynomial approximation are also about 8.90E-03 when the sample size is larger than 100000. Similar results can be found in Table 6. The means of the delta derived by IPA based on the fourth-order polynomial are close to 6.59E-03, the true delta with respect to $S_1(0)$. (ii) Even though the IS parameter (12) is designed

Table 5: Means, variances, and variance reduction ratios of option prices for MC and IS with the fourth-order polynomial approximation and the true objective function (true price is 8.90E-03).

	Fourth-order polynomial approximation					True objective function				
Size	IS		MC			IS		MC		
	mean	variance	mean	variance	ratio	mean	variance	mean	variance	ratio
1000	8.93E-03	1.15E-07	8.54E-03	1.26E-05	109.4	8.88E-03	1.07E-07	9.20E-03	1.55E-05	145.7
2000	8.90E-03	5.82E-08	8.88E-03	5.76E-06	98.8	8.90E-03	4.51E-08	9.12E-03	8.33E-06	184.8
5000	8.91E-03	2.13E-08	8.94E-03	3.25E-06	152.5	8.89E-03	2.16E-08	8.92E-03	3.42E-06	158.1
10000	8.90E-03	1.26E-08	8.93E-03	1.08E-06	85.6	8.90E-03	1.02E-08	8.87E-03	1.16E-06	113.7
20000	8.89E-03	5.93E-09	8.80E-03	6.70E-07	113.0	8.90E-03	5.90E-09	8.87E-03	6.57E-07	111.3
50000	8.90E-03	2.03E-09	8.84E-03	2.65E-07	130.6	8.90E-03	1.94E-09	8.97E-03	2.92E-07	150.3
100000	8.90E-03	1.22E-09	8.92E-03	1.36E-07	111.3	8.90E-03	1.26E-09	8.92E-03	1.45E-07	114.8
200000	8.90E-03	5.05E-10	8.91E-03	5.91E-08	117.0	8.90E-03	4.82E-10	8.92E-03	7.35E-08	152.5
500000	8.90E-03	2.01E-10	8.91E-03	2.73E-08	135.7	8.90E-03	2.25E-10	8.92E-03	2.68E-08	119.2

Table 6: Means, variances, and variance reduction ratios of deltas with respect to $S_1(0)$ for MC and IS with the fourth-order polynomial approximation and the true objective function (true delta with respect to $S_1(0)$ is 6.59E-03).

	Delta for polynomial approximation					Delta for true objective function				
Size	IS		MC			IS		MC		
	mean	variance	mean	variance	ratio	mean	variance	mean	variance	ratio
1000	6.55E-03	1.16E-07	6.41E-03	3.46E-06	29.8	6.61E-03	1.08E-07	6.81E-03	3.66E-06	33.8
2000	6.59E-03	5.88E-08	6.54E-03	1.72E-06	29.3	6.57E-03	5.63E-08	6.64E-03	2.21E-06	39.3
5000	6.58E-03	2.36E-08	6.68E-03	9.35E-07	39.6	6.59E-03	2.24E-08	6.58E-03	9.52E-07	42.6
10000	6.59E-03	1.28E-08	6.59E-03	3.68E-07	28.8	6.59E-03	1.27E-08	6.56E-03	3.62E-07	28.6
20000	6.58E-03	5.91E-09	6.56E-03	1.72E-07	29.1	6.59E-03	6.70E-09	6.57E-03	1.84E-07	27.5
50000	6.59E-03	2.08E-09	6.58E-03	7.40E-08	35.5	6.59E-03	2.03E-09	6.63E-03	7.64E-08	37.7
100000	6.59E-03	1.24E-09	6.61E-03	4.07E-08	32.7	6.59E-03	1.13E-09	6.59E-03	4.04E-08	35.8
200000	6.60E-03	4.49E-10	6.60E-03	2.11E-08	47.1	6.59E-03	6.04E-10	6.61E-03	1.96E-08	32.4
500000	6.59E-03	2.30E-10	6.60E-03	8.38E-09	36.4	6.59E-03	2.39E-10	6.59E-03	7.09E-09	29.6

for the polynomial function, it also works for the original exponential objective function. For example, in Table 5, when the sample size is 200000, the variance reduction ratio for the fourth-order polynomial approximation is 117.0, while the variance reduction ratio for the true objective function is 152.5. (iii) The proposed IS parameter (12) works well for both the price estimation (objective estimation) and delta estimation (gradient estimation). As seen in the tables, when the sample size is small (1000 to 20000), the IS method has good accuracy but the MC does not. The variance reduction ratios for the price estimation and delta estimation are about 100 and 30, respectively.

Next, we change K from 58 to 70, so that $\{g(X) \geq \log K\}$ becomes rarer, and we use the efficiency ratios defined in (6) and (7) to measure the efficiency of the price and delta estimation. Let the sample size be 10^7. Then we obtain Tables 7 and 8. From the tables, we can see that for both the approximated fourth-order

Table 7: Efficiency ratios for IS and MC in price estimation (sample size 10^7).

	K	58	60	62	64	66	68	70
IS	Approximated	1.14E+00	1.27E+00	1.71E+00	2.20E+00	2.28E+00	2.99E+00	3.63E+00
	True objective	1.14E+00	1.27E+00	1.84E+00	1.94E+00	2.28E+00	3.04E+00	3.62E+00
MC	Approximated	2.60E+01	1.73E+02	2.11E+03	4.62E+04	1.62E+06	8.42E+07	6.38E+10
	True objective	2.60E+01	1.73E+02	2.14E+03	4.63E+04	1.62E+06	8.71E+07	6.37E+10

Table 8: Efficiency ratios for IS and MC in delta estimation (sample size 10^7).

	K	58	60	62	64	66	68	70
IS	approximated	5.49E-01	1.18E+00	2.74E+00	5.04E+00	7.82E+00	1.27E+01	1.83E+01
	True objective	5.49E-01	1.18E+00	2.82E+00	4.67E+00	7.81E+00	1.36E+01	1.83E+01
MC	approximated	4.74E+00	4.94E+01	9.66E+02	2.90E+04	1.27E+06	8.51E+07	1.64E+10
	True objective	4.74E+00	4.94E+01	1.16E+03	2.68E+04	1.27E+06	8.11E+07	1.63E+10

polynomial function and the original exponential objective function, as the rare-event parameter becomes larger, the efficiency ratios when using IS do not grow significantly, whereas the efficiency ratios grow very fast when using classical MC.

4 CONCLUSION

This paper studies importance sampling for stochastic gradient estimation involving rare events. We develop theoretical results to show that, for a class of problems, an efficient IS mixture sampler for estimating the probability of the underlying rare event will also be efficient for gradient estimations of a wide class of performance measures defined on the same rare-event set. Several simulation examples, including a financial derivatives example, are provided and support the theory.

ACKNOWLEDGMENTS

We gratefully acknowledge support from the National Science Foundation under grants CAREER CMMI-1834710, IIS-1849280 and IIS-2123684, from the Air Force Office of Scientific Research under grant FA95502010211, and from the National Natural Science Foundation of China under grants 72171060 and 71801148.

REFERENCES

Bai, Y., Z. Huang, H. Lam, and D. Zhao. 2022. "Rare-Event Simulation for Neural Network and Random Forest Predictors". *ACM Transactions on Modeling and Computer Simulation* Forthcoming.

Blanchet, J. H., and H. Lam. 2012. "State-Dependent Importance Sampling for Rare-Event Simulation: An Overview and Recent Advances". *Surveys in Operations Research and Management Science* 17(1):38–59.

Bucklew, J. A. 2004. *Introduction to Rare Event Simulation*. New York, NY: Springer Science & Business Media.

Driggs, D., M. J. Ehrhardt, and C.-B. Schönlieb. 2022. "Accelerating Variance-Reduced Stochastic Gradient Methods". *Mathematical Programming* 191(2):671–715.

Fox, B. L., and P. W. Glynn. 1989. "Replication Schemes For Limiting Expectations". *Probability in the Engineering and Informational Sciences* 3(3):299–318.

Fu, M. C. 2006. "Gradient estimation". In *Handbooks in Operations Research and Management Science*, edited by S. G. Henderson and B. L. Nelson, Volume 13, Chapter 19, 575–616. Amsterdam: Springer.

Fu, M. C., and J. Q. Hu. 1997. *Conditional Monte Carlo: Gradient Estimation and Optimization Applications*. Norwell, MA: Kluwer Academic Publishers.

Glasserman, P. 1988. "Performance Continuity and Differentiability in Monte Carlo Optimization". In *Proceedings of the 1988 Winter Simulation Conference*, edited by M. Abrams, P. Haigh, and J. Comfort, 518–524. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Glynn, P. 1989. "Optimization of Stochastic Systems via Simulation". In *Proceedings of the 1989 Winter Simulation Conference*, edited by E. A. MacNair, K. J. Musselman, and P. Heidelberger, 90–105. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Glynn, P. 1990. "Likelihood Ratio Gradient Estimation for Stochastic Systems". *Communications of the ACM* 33(10):75–84.

He, S., G. Jiang, H. Lam, and M. C. Fu. 2022. "Adaptive Importance Sampling for Efficient Stochastic Root Finding and Quantile Estimation". working paper, https://arxiv.org/abs/2102.10631.

Heidergott, B., F. Vazquez-Abad, G. Pflug, and T. Farenhorst-Yuan. 2010. "Gradient Estimation for Discrete-Event Systems by Measure-Valued Differentiation". *ACM Transactions on Modeling and Computer Simulation* 20(1):No. 5.

Ho, Y. C., and X. Cao. 1983. "Perturbation Analysis and Optimization of Queueing Networks". *Journal of Optimization Theory and Applications* 40:559–582.

Johnson, R., and T. Zhang. 2013. "Accelerating Stochastic Gradient Descent Using Predictive Variance Reduction". In *Advances in Neural Information Processing Systems*, 315–323. Red Hook, NY, USA: Curran Associates Inc.

Juneja, S., and P. Shahabuddin. 2006. "Rare-Event Simulation Techniques: An Introduction and Recent Advances". In *Handbooks in Operations Research and Management Science*, edited by S. G. Henderson and B. L. Nelson, Volume 13, Chapter 11, 291 – 350. Amsterdam, The Netherlands: Elsevier.

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

Liu, G., and L. J. Hong. 2011. "Kernel Estimation of the Greeks for Options with Discontinuous Payoffs". *Operations Research* 59(1):96–108.

Nakayama, M. K. 1995. "Asymptotics of Likelihood Ratio Derivative Estimators in Simulations of Highly Reliable Markovian Systems". *Management Science* 41(3):524–554.

Peng, Y. J., M. C. Fu, J. Q. Hu, and B. Heidergott. 2018. "A New Unbiased Stochastic Derivative Estimator for Discontinuous Sample Performances with Structural Parameters". *Operations Research* 66(2):487–499.

Pflug, G. C. 1989. "Sampling Derivatives of Probabilities". *Computing* 42:315–328.

Reiman, M. I., and A. Weiss. 1989. "Sensitivity Analysis for Simulations via Likelihood Ratios". *Operations Research* 37(5):830–844.

Rubinstein, R. Y. 1989. "Sensitivity Analysis and Performance Extrapolation for Computer Simulation Models". *Operations Research* 37(1):72–81.

Wang, C., X. Chen, A. J. Smola, and E. P. Xing. 2013. "Variance Reduction for Stochastic Gradient Optimization". In *Advances in Neural Information Processing Systems*, edited by C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, Volume 26: Curran Associates, Inc.

Wang, Y., M. C. Fu, and S. I. Marcus. 2012. "A New Stochastic Derivative Estimator for Discontinuous Payoff Functions with Application to Financial Derivatives". *Operations Research* 60(2):447–460.

Xiao, L., and T. Zhang. 2014. "A Proximal Stochastic Gradient Method with Progressive Variance Reduction". *SIAM Journal on Optimization* 24(4):2057–2075.

Zhao, P., and T. Zhang. 2015, 07–09 Jul. "Stochastic Optimization with Importance Sampling for Regularized Loss Minimization". In *Proceedings of the 32nd International Conference on Machine Learning*, edited by F. Bach and D. Blei, Volume 37, 1–9. Lille, France: PMLR.

AUTHOR BIOGRAPHIES

YUANLU BAI is a PhD candidate in the Department of Industrial Engineering and Operations Research at Columbia University. She received an M.S. in operations research from Columbia University, and a B.S. in statistics as well as a B. Ec in economics from Peking University. Her research interest lies in uncertainty quantification and rare-event simulation. Her email address is yb2436@columbia.edu.

SHENGYI HE is a PhD student in the Department of Industrial Engineering and Operations Research at Columbia University. He received his B.S. degree in statistics from Peking University in 2019. His research interests include variance reduction and uncertainty quantification via stochastic and robust optimization. His email address is sh3972@columbia.edu.

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.

GUANGXIN JIANG is a professor in the School of Management at Harbin Institute of Technology. He earned his Ph.D. in Applied Mathematics from Tongji University. His research interests include stochastic models and simulation, machine learning, financial engineering and risk management, FinTech, etc. His email address is gxjiang@hit.edu.cn.

MICHAEL C. FU holds the Smith Chair of Management Science in the Robert H. Smith School of Business, with a joint appointment in the Institute for Systems Research and affiliate faculty appointment in the Department of Electrical and Computer Engineering, all at the University of Maryland, College Park. His research interests include stochastic gradient estimation, simulation optimization, and applied probability. He served as WSC2011 Program Chair and received the INFORMS Simulation Society's Distinguished Service Award in 2018. He is a Fellow of INFORMS and IEEE. His e-mail addresses is mfu@umd.edu.