Proceedings of the 2016 Winter Simulation Conference
T. M. K. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick, eds.

THE EMPIRICAL LIKELIHOOD APPROACH TO SIMULATION INPUT UNCERTAINTY

Henry Lam

Department of Industrial and Operations Engineering
University of Michigan
1205 Beal Ave.
Ann Arbor, MI 48109, USA

Huajie Qian

Department of Mathematics
University of Michigan
2074 East Hall
530 Church Street
Ann Arbor, MI 48109, USA

ABSTRACT

We study the empirical likelihood method in constructing statistically accurate confidence bounds for stochastic simulation under nonparametric input uncertainty. The approach is based on positing a pair of distributionally robust optimization, with a suitably averaged divergence constraint over the uncertain input distributions, and calibrated with a χ^2-quantile to provide asymptotic coverage guarantees. We present the theory giving rise to the constraint and the calibration. We also analyze the performance of our stochastic optimization algorithm. We numerically compare our approach with existing standard methods such as the bootstrap.

1 INTRODUCTION

Stochastic simulation often relies on input distributions that are not fully known but only observed via a finite sample of input data. The quantification of the errors in performance analyses due to the misspecification of input models is known as input uncertainty. This paper focuses particularly on situations where there are no specific assumptions on the parametric form of the input models, or nonparametric input uncertainty. One major goal in this context is to construct confidence interval (CI) for the target performance measure that accounts for input model errors on top of simulation errors.

This paper proposes an optimization-based method to construct such a CI. The upper and lower bounds of the CI are obtained as the optimal values of a pair of optimization programs posited over the space of input probability distributions. These programs can be interpreted as finding the upper and lower worst-case performance measures, subject to a constraint on the unknown input distributions within a neighborhood of the empirical distributions. We show that, by choosing the right size of this neighborhood and the appropriate nonparametric distance to measure its size, these worst-case optimizations lead to asymptotically exact confidence bounds for covering the true performance measure. We develop these statistical guarantees by using the empirical likelihood (EL) method, which can be viewed as a nonparametric analog of the maximum likelihood theory.

Our approach resembles, but also should be contrasted with, the literature of distributionally robust optimization (DRO) (e.g., Delage and Ye 2010, Ben-Tal et al. 2013, Goh and Sim 2010). This literature aims to find worst-case stochastic performance measures (or optimizes decisions over the worst-case) subject to so-called uncertainty sets that represent the information or uncertainty about an underlying probability distribution. One common constraint is a nonparametric neighborhood ball around a baseline model, typically corresponding to the modeler's best guess of the truth, measured by some nonparametric distance such as ϕ-divergence (Pardo 2005). When this ball contains the true distribution, the optimal values of the worst-case optimizations will cover the true measure. Thus, it is argued that one should calibrate the ball size by estimating the divergence between the data and the baseline distribution, and consequently translate the confidence of this estimation into a confidence bound for the performance measure. In practice, this

implies calibration using a goodness-of-fit χ^2-quantile, with the degree of freedom corresponding to the effective number of categories in a discrete distribution, or divergence estimation that involves estimating a functional of density. These methods could run into the statistical challenges in divergence estimation, and also potentially a loss of coverage accuracy in translating the uncertainty set into a CI.

One main contribution of this paper is thus a new way of interpreting these DROs via the EL method. Through this connection we locate the precise nonparametric distance one should use in constructing valid CIs under nonparametric input uncertainty. Our distance deviates from those in the literature as it consists of a weighted average of a collection of divergences, each applied on an independent input model needed in the simulation. We also show that the size of the nonparametric neighborhood can be taken as a χ^2-quantile, with degree of freedom exactly one, independent of the number of input models and the continuity of these random variates. This is in sharp contrast with the proposed calibration methods in the DRO literature.

In general, the optimization programs we impose is simulation-based. As another contribution, we investigate the properties of mirror descent stochastic approximation (MDSA) (Nemirovski et al. 2009) used to locally solve these programs. This simulation-based optimization framework provides an alternate approach to quantifying nonparametric input uncertainty compared to the current major technique of bootstrap resampling. The latter is a sampling approach that repeatedly generates new empirical distributions to drive the simulation runs. On a high level, our approach trades the computational load in resampling with the optimization routine needed in the iteration of the SA algorithm. We investigate this tradeoff, and compare the performances of the obtained CIs and their vulnerability to the simulation noise.

The rest of this paper is organized as follows. Section 2 lays out our optimization programs and explains them using the EL method. Section 3 presents and analyzes our MDSA algorithm. Section 4 show some numerical results and comparison with the bootstrap.

2 EMPIRICAL-LIKELIHOOD-BASED CONFIDENCE INTERVAL

We consider a performance measure in the form

$$Z(P_1,\ldots,P_m)=\mathbb{E}_{P_1,\ldots,P_m}\left[h(\mathbf{X}_1,\ldots,\mathbf{X}_m)\right], \tag{1}$$

where each $\mathbf{X}_i=(X_i(1),\ldots,X_i(T_i))$ is a sequence of T_i i.i.d. random variables under an independent input model (or distribution) P_i, and T_i is a deterministic run length. The function h mapping from $\mathfrak{X}_1^{T_1}\times\cdots\times\mathfrak{X}_m^{T_m}$ to $\mathbb{R}$ is assumed computable given the inputs $\mathbf{X}_i$'s.

Our premise is that each P_i is unknown but n_i i.i.d. data $X_{i,1},\ldots,X_{i,n_i}$ are available. The true value of (1) is therefore unknown even under abundant simulation runs. Our goal is to find a $(1-\alpha)$ level CI for the true performance measure.

Our approach is the following. For model i, we consider the probability simplex of the weights $\mathbf{w}_i=(w_{i,1},\ldots,w_{i,n_i})$ over the support set $\{X_{i,1},\ldots,X_{i,n_i}\}$. We consider the pair of optimization programs

$$\begin{aligned}\mathscr{L}_\alpha/\mathscr{U}_\alpha:=\min/\max\quad & Z(\mathbf{w}_1,\ldots,\mathbf{w}_m)\\ \text{subject to}\quad & -2\sum_{i=1}^{m}\sum_{j=1}^{n_i}\log n_i w_{i,j}\le\mathcal{X}^2_{1,1-\alpha}\\ & \sum_{j=1}^{n_i}w_{i,j}=1,\ \text{for all } i\\ & w_{i,j}\ge 0,\ \text{for all } i,j\end{aligned} \tag{2}$$

where $\mathcal{X}^2_{1,1-\alpha}$ is the $1-\alpha$ quantile of the chi-square distribution with degree of freedom one, and each $\mathbf{w}_i$ in $Z(\mathbf{w}_1,\ldots,\mathbf{w}_m)$ should be interpreted as the distribution putting weight $w_{i,j}$ on $X_{i,j}$, for $j=1,\ldots,n_i$.

The formulation (2) can be interpreted as the worst-case optimizations over the m independent input distributions subject to a weighted-averaged divergence. Note that the quantity $-(1/n_i)\sum_{j=1}^{n_i}\log n_i w_{i,j}$ is

the Burg-entropy divergence (Ben-Tal et al. 2013) between the probability weights $\mathbf{w}_i$ and the uniform weights. Thus, letting $n = \sum_{i=1}^m n_i$ be the total sample size, we have

$$-\frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\log n_i w_{i,j} = \sum_{i=1}^{m}\frac{n_i}{n}\left(-\frac{1}{n_i}\sum_{j=1}^{n_i}\log n_i w_{i,j}\right)$$

which can be viewed as a weighted average of the individual Burg-entropy divergences imposed on different input models, each weight being n_i/n. The first constraint in (2) imposes a bound $\mathcal{X}^2_{1,1-\alpha}/(2n)$ on this averaged divergence.

2.1 Empirical Likelihood Theory for Sum of Means

We justify the formulation (2) using the EL method. First proposed by Owen (1988), this method can be viewed as a nonparametric counterpart of maximum likelihood theory. Analogous to Wilks' Theorem (Cox and Hinkley 1979) in maximum likelihood theory that states the convergence of the so-called logarithmic likelihood ratio to a χ^2-distribution, the nonparametric profile likelihood ratio in EL converges similarly. This will be the key to obtaining our formulation (2).

We describe the EL method. Given m independent sets of data $\{X_{i,1},\ldots,X_{i,n_i}\}$, we define the nonparametric likelihood, in terms of the probability weights $\mathbf{w}_i$ for the i-th input model, to be $\prod_{j=1}^{n_i} w_{i,j}$. The multi-sample likelihood is $\prod_{i=1}^m\prod_{j=1}^{n_i} w_{i,j}$. It can be shown, by a simple convexity argument, that uniform weights $w_{i,j} = 1/n_i$ for each model maximize $\prod_{i=1}^m (1/n_i)^{n_i}$. Moreover, uniform weights still maximize even if one allows putting weights outside the support of data, in which case $\sum_{j=1}^{n_i} w_{i,j} < 1$ for some i, making $\prod_{i=1}^m\prod_{j=1}^{n_i} w_{i,j}$ even smaller. Therefore, $\prod_{i=1}^m (1/n_i)^{n_i}$ can be viewed as the nonparametric maximum likelihood estimate.

To proceed, we need to define a parameter of interest that is determined by the input models. Inference about this parameter is carried out based on the so-called profile nonparametric likelihood ratio, which is the maximum likelihood ratio between all weights that empirically produce a given value of the parameter, and the uniform weights (i.e. the MLE). In our case the parameter of interest is the performance measure $Z(P_1,\ldots,P_m)$, which is however possibly nonlinear in P_i's and hence not easy to work with. Fortunately, it turns out that there isn't much loss to deal with the special case $h(\mathbf{X}_1,\ldots,\mathbf{X}_m) = \sum_{i=1}^m h_i(X_i(1))$ for some $h_i : \mathbb{R}\to\mathbb{R}$ and $T_i = 1$ for all i, i.e. the parameter of interest is simply the sum of means of random variables $h_i(X_i(1))$. We will focus on this case for now. To simplify further, let us consider estimating $\sum_{i=1}^m \mathbb{E}X_i$, where for convenience we write X_i in place of $X_i(1)$. The profile nonparametric likelihood ratio is defined as

$$R(\mu) = \max\left\{\prod_{i=1}^{m}\prod_{j=1}^{n_i} n_i w_{i,j}\,\middle|\,\sum_{i=1}^{m}\sum_{j=1}^{n_i} w_{i,j}X_{i,j} = \mu,\ \sum_{j=1}^{n_i} w_{i,j} = 1, w_{i,j}\ge 0,\ \text{for all } i,j\right\}, \tag{3}$$

and is defined to be 0 if the optimization problem in (3) is infeasible.

Before getting into details of the asymptotic theory, we note that usually the logarithmic profile nonparametric likelihood ratio at the true value, i.e. $-2\log R(\sum_{i=1}^m \mathbb{E}X_i(1))$, asymptotically follows some chi-square distribution whose degree of freedom is equal to the number of non-probability-simplex constraints, e.g. $\sum_{i=1}^m\sum_{j=1}^{n_i} w_{i,j}X_{i,j} = \mu$ in (3). As another interpretation, treating the weights $\mathbf{w}_i$ as parameters, the degree of freedom is the difference in dimensions of the full and constrained parameter space, and typically d constraints result in a loss of d dimensions of the parameter space.

Now we state our theorem.

Theorem 1 Let X_i be a random variable distributed under P_i. Assume $0 < \sum_{i=1}^m \mathrm{Var}(X_i) < \infty$, and $\min_{i\in I} n_i \ge c\sum_{i\in I} n_i$ always holds for some constant $c > 0$, where $I = \{i|\mathrm{Var}(X_i) > 0\}$. Then $-2\log R(\sum_{i=1}^m \mathbb{E}X_i)$ converges in distribution to $\mathcal{X}^2_1$, the chi-squared distribution with degree of freedom one, as $n_i\to\infty$ for $i\in I$.

In this theorem only sample sizes of those having positive variance are required to grow to infinity. To see the reason for this, note that data of inputs with zero variance are always equal to the true mean so we

can always assign them the uniform weights in (3). Other than these zero-variance inputs, the condition $\min_{i\in I} n_i \geq c\sum_{i\in I} n_i$ basically forces the sample sizes to grow at the same rate. Theorem 1 is a multi-sample generalization of the well known empirical likelihood theorem for single-sample mean, which can be found in Chapter 2 of Owen (2001). We state it as a special case where $m = 1$.

Theorem 2 Let $Y_1, Y_2, \ldots, Y_n$ be i.i.d. random variables distributed under some distribution P, $0 < \mathrm{Var}(Y_1) < \infty$. Then $-2\log R(\mathbb{E}Y_1)$ converges in distribution to $\mathcal{X}_1^2$, as $n \to \infty$. The function $R(\cdot)$ here is the same as that in (3) with $m = 1, n_1 = n, X_{1,j} = Y_j$.

2.2 Empirical-likelihood-based Confidence Interval

We discuss how to construct confidence interval for the quantity $Z(P_1, \ldots, P_m)$ based on the EL theory presented in the last section, and thus justify the validity of the formulation (2). We will first study the linear output case, and then discuss the general performance measure in (1).

As pointed out in the last section, linear output takes the form $h(\mathbf{X}_1, \ldots, \mathbf{X}_m) = \sum_{i=1}^m h_i(X_i)$, where X_i is distributed under P_i. For output of this particular form, Theorem 1 implies that the optimization pair (2) gives an asymptotically correct confidence interval with coverage probability $1-\alpha$, namely

Theorem 3 Assume $0 < \sum_{i=1}^m \mathrm{Var}(h_i(X_i)) < \infty$, and $\min_{i\in I} n_i \geq c\sum_{i\in I} n_i$ for some constant $c > 0$, where $I = \{i | \mathrm{Var}(h_i(X_i)) > 0\}$. Then

$$P\left(\mathscr{L}_\alpha \leq \sum_{i=1}^m \mathbb{E}h_i(X_i) \leq \mathscr{U}_\alpha\right) \to 1-\alpha, \text{ as } n_i \to \infty \text{ for } i \in I,$$

where

$$\begin{aligned} \mathscr{L}_\alpha / \mathscr{U}_\alpha := \min / \max \quad & \sum_{i=1}^m \sum_{j=1}^{n_i} w_{i,j} h_i(X_{i,j}) \\ \text{subject to} \quad & -2\sum_{i=1}^m \sum_{j=1}^{n_i} \log n_i w_{i,j} \leq \mathcal{X}^2_{1,1-\alpha} \\ & \sum_{j=1}^{n_i} w_{i,j} = 1, \text{ for all } i \\ & w_{i,j} \geq 0, \text{ for all } i, j \end{aligned} \tag{4}$$

and $\mathcal{X}^2_{1,1-\alpha}$ is the $1-\alpha$ quantile of $\mathcal{X}_1^2$.

Proof. By applying Theorem 1 to $h_i(X_{i,j})$, we know that the set $\{\mu \in \mathbb{R} | -2\log R(\mu) \leq \mathcal{X}^2_{1,1-\alpha}\}$ contains the true sum $\sum_{i=1}^m \mathbb{E}h_i(X_i)$ with probability $1-\alpha$ asymptotically. Note that this set can be identified as

$$\mathcal{U} = \left\{ \sum_{i=1}^m \sum_{j=1}^{n_i} w_{i,j} h_i(X_{i,j}) \middle| -2\sum_{i=1}^m \sum_{j=1}^{n_i} \log n_i w_{i,j} \leq \mathcal{X}^2_{1,1-\alpha}, \sum_{j=1}^{n_i} w_{i,j} = 1, w_{i,j} \geq 0 \right\}.$$

It is obvious that $\mathscr{L}_\alpha / \mathscr{U}_\alpha = \min / \max\{\mu | \mu \in \mathcal{U}\}$. So if the set $\mathcal{U}$ is convex, then $\mathcal{U} = [\mathscr{L}_\alpha, \mathscr{U}_\alpha]$, and we conclude the theorem. To show convexity, it's enough to notice that the feasible set of (4) is convex, and the objective is linear in $w_{i,j}$. □

We now extend our result to the general performance measure (1). We show that, by decomposing $Z(\mathbf{w}_1, \ldots, \mathbf{w}_m)$ into linear and nonlinear parts via a Taylor-like expansion, applying Theorem 3 to the linear part, and controlling the magnitude of the nonlinear part, we can construct a CI that has an asymptotic guarantee similar to Theorem 3. We assume that

Assumption 1 $\sum_{i=1}^m \mathrm{Var}(G_i(X_i)) > 0$, where $G_i(x) = \sum_{j=1}^{T_i} \mathbb{E}_{P_1,\ldots,P_m}[h(\mathbf{X}_1, \ldots, \mathbf{X}_m) | X_i(j) = x]$.

Assumption 2 Denote the index $I_i = (I_i(1), \ldots, I_i(T_i)) \in \{1, 2, \ldots, T_i\}^{T_i}$, and $\mathbf{X}_{i,I_i} = (X_i(I_i(1)), \ldots, X_i(I_i(T_i)))$. Assume that there exists some positive integer k such that $h(\mathbf{X}_{1,I_1}, \ldots, \mathbf{X}_{m,I_m})$ has finite $2k$-th moment for all possible choices of I_i's.

It is not difficult to see that Assumption 1 is the counterpart of the condition $\sum_{i=1}^m \mathrm{Var}(h_i(X_i)) > 0$ in Theorem 3, which is needed in dealing with the linear part of $Z(\mathbf{w}_1, \ldots, \mathbf{w}_m)$, and Assumption 2 is used for controlling the nonlinear part. We state our result for the general performance measure below.

Theorem 4 If Assumptions 1, 2 hold, and $\min_i n_i \geq c \sum_{i=1}^m n_i$ always holds for some constant $c > 0$, then

$$P\left(\mathscr{L}_\alpha + O_p\left(\frac{1}{n^{\frac{k-1}{k}}}\right) \leq Z(P_1, \ldots, P_m) \leq \mathscr{U}_\alpha + O_p\left(\frac{1}{n^{\frac{k-1}{k}}}\right)\right) \to 1 - \alpha, \text{ as } n_i \to \infty,$$

where $n = \sum_{i=1}^m n_i$ and $\mathscr{L}_\alpha, \mathscr{U}_\alpha$ are defined as in (2).

This theorem shows that if at least sixth moments ($k \geq 3$) of the function h are finite, then the difference $O_p(1/n^{\frac{k-1}{k}})$ is asymptotically negligible compared with $O_p(1/\sqrt{n})$, the length of the CI. This holds trivially if, for instance, h is a bounded function.

3 MIRROR DESCENT STOCHASTIC APPROXIMATION

This section focuses on solving optimization problem (2), using the mirror descent (MD) algorithm. The MD algorithm was first proposed by Nemirovsky et al. (1982) for deterministic convex optimization in general normed space, motivated by the challenge that the standard gradient descent in the Euclidean space may not generally make sense because the primal space, where the solution lies on, can be different from its dual, where the gradient is defined on. It resolves the issue by mapping the solution to the dual space and conducts gradient descent therein. For optimizations in $\mathbb{R}^n$, one does not necessarily need to use MD since the primal and the dual space are the same. However, with a given set of constraints, a judicious choice of a primal-dual map can result in iteration subroutine that is computationally efficient. Such observations extend to the setting where gradient can only be simulated. In this case, the algorithm becomes a constrained SA procedure, and leads to MDSA.

Here we outline the MDSA procedure for the following generic version of problem (2) with $\eta > 0$

$$\begin{aligned} \min \quad & Z(\mathbf{w}_1, \ldots, \mathbf{w}_m) \\ \text{subject to} \quad & -2\sum_{i=1}^m \sum_{j=1}^{n_i} \log n_i w_{i,j} \leq \eta \\ & \sum_{j=1}^{n_i} w_{i,j} = 1, \text{ for all } i \\ & w_{i,j} \geq 0, \text{ for all } i, j \end{aligned} \tag{5}$$

The feasible set here, denoted by $\mathscr{A}$, is convex and satisfies $\mathscr{A} \subset \prod_{i=1}^m \mathscr{P}^{n_i} \subset \mathbb{R}^n$, where $\mathscr{P}^{n_i}$ denotes the probability simplex associated with a support set of size n_i, and $n = \sum_{i=1}^m n_i$. The norm on $\mathbb{R}^n$ is chosen to be the 1-norm $\|(\mathbf{w}_1, \ldots, \mathbf{w}_m)\|_1 = \sum_{i=1}^m \sum_{j=1}^{n_i} |w_{i,j}|$. First we take the entropy distance generating function ω on $\prod_{i=1}^m \mathscr{P}^{n_i}$ (used in the so-called entropic descent algorithm (Beck and Teboulle 2003))

$$\omega(\mathbf{w}_1, \ldots, \mathbf{w}_m) = \sum_{i=1}^m \sum_{j=1}^{n_i} w_{i,j} \log w_{i,j}.$$

Strong convexity of ω is guaranteed by the following proposition, whose proof is similar to that of the special case $m = 1$ in Nemirovski et al. (2009).

Proposition 1 $\omega(\mathbf{w}_1,\ldots,\mathbf{w}_m)$ is strongly convex with parameter $1/m$ w.r.t. 1-norm in the relative interior of $\prod_{i=1}^m \mathscr{P}^{n_i}$, i.e.

$$\omega(\mathbf{u}_1,\ldots,\mathbf{u}_m)-\omega(\mathbf{w}_1,\ldots,\mathbf{w}_m)\geq\sum_{i=1}^m\sum_{j=1}^{n_i}\left.\frac{\partial\omega}{\partial w_{i,j}}\right|_{(\mathbf{w}_1,\ldots,\mathbf{w}_m)}(u_{i,j}-w_{i,j})+\frac{1}{2m}\left(\sum_{i=1}^m\|\mathbf{u}_i-\mathbf{w}_i\|_1\right)^2.$$

Corresponding to ω is the prox-function responsible for projecting the solution back to the primal decision space

$$\begin{aligned}V(\mathbf{u}_1,\ldots,\mathbf{u}_m;\mathbf{w}_1,\ldots,\mathbf{w}_m)&=\omega(\mathbf{u}_1,\ldots,\mathbf{u}_m)-\omega(\mathbf{w}_1,\ldots,\mathbf{w}_m)-\sum_{i=1}^m\sum_{j=1}^{n_i}\left.\frac{\partial\omega}{\partial w_{i,j}}\right|_{(\mathbf{w}_1,\ldots,\mathbf{w}_m)}(u_{i,j}-w_{i,j})\\&=\sum_{i=1}^m\sum_{j=1}^{n_i}w_{i,j}\log\frac{w_{i,j}}{u_{i,j}}.\end{aligned}$$

In each step of MDSA, the current solution $(\mathbf{w}_1^k,\ldots,\mathbf{w}_m^k)$ is updated via the prox-mapping

$$\text{prox}(\mathbf{w}_1^k,\ldots,\mathbf{w}_m^k)=\text{argmin}_{(\mathbf{w}_1,\ldots,\mathbf{w}_m)\in\mathscr{A}}\gamma_k\sum_{i,j}\widehat{\frac{\partial Z}{\partial w_{i,j}}}(w_{i,j}-w_{i,j}^k)+V(\mathbf{w}_1^k,\ldots,\mathbf{w}_m^k;\mathbf{w}_1,\ldots,\mathbf{w}_m),\tag{6}$$

where γ_k is the step size, and $\widehat{\partial Z/\partial w_{i,j}}$ is an estimate for each component of the gradient.

3.1 Gradient Estimation

Here we discuss how to estimate the gradient of the objective $Z(\mathbf{w}_1,\ldots,\mathbf{w}_m)$, which is needed in the prox-mapping (6). It is not hard to see that $Z(\mathbf{w}_1,\ldots,\mathbf{w}_m)$ is a multivariate polynomial in $w_{i,j}$, and hence its gradient is also a polynomial. However, a closer examination reveals that the number of terms in this polynomial grows as fast as n^T, where n is the order of the sample size and T the time horizon. Thus it is generally impossible to compute the gradient by naively summing up all the terms, and simulation is needed. We adopt the proposal by Ghosh and Lam (2015a) and Ghosh and Lam (2015b) of using the directional derivative $\psi_{i,j}=\frac{d}{d\varepsilon}Z(\mathbf{w}_1,\ldots,(1-\varepsilon)\mathbf{w}_i+\varepsilon\mathbf{e}_{i,j},\ldots,\mathbf{w}_m)|_{\varepsilon=0}$ instead of the standard derivative $\partial Z/\partial w_{i,j}$, where $\mathbf{e}_{i,j}$ is the j-th coordinate vector of $\mathbb{R}^{n_i}$. It's shown that substituting $\psi_{i,j}$ in place of $\partial Z/\partial w_{i,j}$ retains all properties needed in the prox-mapping (6), and appealingly $\psi_{i,j}$ is simulable. Here is an extension of their result.

Proposition 2 For any $(\mathbf{w}_1,\ldots,\mathbf{w}_m)\in\prod_{i=1}^m\mathscr{P}^{n_i}$ with each $w_{i,j}>0$, let $\psi_{i,j}=\frac{d}{d\varepsilon}Z(\mathbf{w}_1,\ldots,(1-\varepsilon)\mathbf{w}_i+\varepsilon\mathbf{e}_{i,j},\ldots,\mathbf{w}_m)|_{\varepsilon=0}$, and $\partial Z/\partial w_{i,j}$ be the standard derivative. Then

$$\sum_{i=1}^m\sum_{j=1}^{n_i}\psi_{i,j}(u_{i,j}-w_{i,j})=\sum_{i=1}^m\sum_{j=1}^{n_i}\frac{\partial Z}{\partial w_{i,j}}(u_{i,j}-w_{i,j}),\text{ for any }(\mathbf{u}_1,\ldots,\mathbf{u}_m)\in\prod_{i=1}^m\mathscr{P}^{n_i}$$

and

$$\psi_{i,j}=E_{\mathbf{w}_1,\ldots,\mathbf{w}_m}\left[h(\mathbf{X}_1,\ldots,\mathbf{X}_m)S_{i,j}(\mathbf{X}_i)\right],$$

where

$$S_{i,j}(\mathbf{X}_i)=\sum_{t=1}^{T_i}\frac{\mathbf{1}\left\{X_i(t)=X_{i,j}\right\}}{w_{i,j}}-T_i.$$

This proposition suggests the following unbiased estimator for $\psi_{i,j}$

$$\hat{\psi}_{i,j}=\frac{1}{R}\sum_{r=1}^R h(\mathbf{X}_1^r,\ldots,\mathbf{X}_m^r)S_{i,j}(\mathbf{X}_i^r),\tag{7}$$

where $\mathbf{X}_1^r,\ldots,\mathbf{X}_m^r, r=1,\ldots,R$ are R independent replications of the m input processes generated under weights $(\mathbf{w}_1,\ldots,\mathbf{w}_m)$, and are used simultaneously in all $\hat{\psi}_{i,j}$'s.

3.2 Computing the Prox-mapping

We discuss how to solve the minimization problem in prox-mapping (6) with $\widehat{\partial Z/\partial w_{i,j}}$ replaced by $\hat{\psi}_{i,j}$. Let's work with the generic form

$$\begin{aligned} \min \quad & \sum_{i=1}^{m}\sum_{j=1}^{n_i} \xi_{i,j} w_{i,j} + V(\mathbf{u}_1,\ldots,\mathbf{u}_m;\mathbf{w}_1,\ldots,\mathbf{w}_m) \\ \text{subject to} \quad & -2\sum_{i=1}^{m}\sum_{j=1}^{n_i} \log n_i w_{i,j} \le \eta \\ & \sum_{j=1}^{n_i} w_{i,j} = 1, \text{ for all } i \\ & w_{i,j} \ge 0, \text{ for all } i,j. \end{aligned} \tag{8}$$

where $\xi_{i,j}$'s are put in place of the gradient estimators. Seeing that the decision space of (8) has dimension $\sum_{i=1}^{m} n_i$, which can be significantly larger than the number of functional constraints $m+1$, our strategy is to first solve the Lagrangian dual problem, then use the dual optimal solution to recover the primal optimal solution.

Consider the Lagrangian

$$\begin{aligned} & L(\mathbf{w}_1,\ldots,\mathbf{w}_m,\boldsymbol{\lambda},\beta) \\ =& \sum_{i=1}^{m}\left[\sum_{j=1}^{n_i}\left(\xi_{i,j}w_{i,j} + w_{i,j}\log\frac{w_{i,j}}{u_{i,j}}\right) + \lambda_i\left(\sum_{j=1}^{n_i} w_{i,j} - 1\right)\right] - \beta\left(2\sum_{i=1}^{m}\sum_{j=1}^{n_i}\log n_i w_{i,j} + \eta\right) \end{aligned}$$

The dual function defined for $\boldsymbol{\lambda}\in\mathbb{R}^m, \beta\ge 0$ is

$$\begin{aligned} g(\boldsymbol{\lambda},\beta) &= \min_{\mathbf{w}_1,\ldots,\mathbf{w}_m\ge 0} L(\mathbf{w}_1,\ldots,\mathbf{w}_m,\boldsymbol{\lambda},\beta) \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n_i}\min_{w_{i,j}\ge 0}\left((\xi_{i,j}+\lambda_i)w_{i,j} + w_{i,j}\log\frac{w_{i,j}}{u_{i,j}} - 2\beta\log n_i w_{i,j}\right) - \sum_{i=1}^{m}\lambda_i - \beta\eta \qquad (9) \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left((\xi_{i,j}+\lambda_i)\tilde{w}_{i,j} + \tilde{w}_{i,j}\log\frac{\tilde{w}_{i,j}}{u_{i,j}} - 2\beta\log n_i\tilde{w}_{i,j}\right) - \sum_{i=1}^{m}\lambda_i - \beta\eta, \qquad (10) \end{aligned}$$

where $\tilde{w}_{i,j}$ is the minimizer of the inner optimization in (9). This inner optimization can be solved efficiently using Newton's method. In other words, $g(\boldsymbol{\lambda},\beta) = L(\tilde{\mathbf{w}}_1,\ldots,\tilde{\mathbf{w}}_m,\boldsymbol{\lambda},\beta)$ can be easily evaluated. Moreover, it can be shown that $\tilde{w}_{i,j}$ is always strictly positive, hence $\partial L/\partial w_{i,j}\big|_{w_{i,j}=\tilde{w}_{i,j}} = 0$. By this observation and the chain rule, the first derivatives of the dual function are

$$\frac{\partial g}{\partial\lambda_i} = \sum_{j=1}^{n_i}\tilde{w}_{i,j} - 1, \frac{\partial g}{\partial\beta} = -2\sum_{i=1}^{m}\sum_{j=1}^{n_i}\log n_i\tilde{w}_{i,j} - \eta. \tag{11}$$

To obtain the second derivatives, note that $\tilde{w}_{i,j}$ is determined by $\xi_{i,j}+\lambda_i+\log\frac{\tilde{w}_{i,j}}{\mu_{i,j}}+1-\frac{2\beta}{\tilde{w}_{i,j}} = 0$. Implicit differentiation gives

$$\frac{\partial\tilde{w}_{i,j}}{\partial\lambda_i} = \frac{-\tilde{w}_{i,j}^2}{\tilde{w}_{i,j}+2\beta}, \ \frac{\partial\tilde{w}_{i,j}}{\partial\lambda_s} = 0, \text{ for } s\ne i, \ \frac{\partial\tilde{w}_{i,j}}{\partial\beta} = \frac{2\tilde{w}_{i,j}}{\tilde{w}_{i,j}+2\beta}.$$

So the second derivatives are

$$\begin{aligned} \frac{\partial^2 g}{\partial \lambda_i^2} &= \sum_{j=1}^{n_i} \frac{-\tilde{w}_{i,j}^2}{\tilde{w}_{i,j}+2\beta}, \qquad \frac{\partial^2 g}{\partial \beta^2} = \sum_{i=1}^{m}\sum_{j=1}^{n_i} \frac{-4}{\tilde{w}_{i,j}+2\beta}, \\ \frac{\partial^2 g}{\partial \lambda_i \partial \lambda_s} &= 0 \text{ for } s \neq i, \qquad \frac{\partial^2 g}{\partial \lambda_i \partial \beta} = \sum_{j=1}^{n_i} \frac{2\tilde{w}_{i,j}}{\tilde{w}_{i,j}+2\beta}. \end{aligned} \tag{12}$$

Therefore we have shown in (10), (11) and (12) that the dual function $g(\boldsymbol{\lambda}, \beta)$, its gradient and its Hessian can all be efficiently evaluated. This allows us to solve the dual problem

$$\begin{aligned} \max_{\boldsymbol{\lambda},\beta} \quad & g(\boldsymbol{\lambda}, \beta) \\ \text{subject to} \quad & \boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m) \in \mathbb{R}^m, \beta \geq 0 \end{aligned} \tag{13}$$

efficiently using, e.g., the Interior Point method. The following proposition helps justify our proposal of solving the dual (13) instead.

Proposition 3 There is a unique optimal solution $(\lambda_1^*, \ldots, \lambda_m^*, \beta^*)$ to the dual (13). Moreover, the primal optimal solution $w_{i,j}^*$ to (8) can be obtained by solving

$$\xi_{i,j} + \lambda_i^* + \log \frac{w_{i,j}^*}{u_{i,j}} + 1 - \frac{2\beta^*}{w_{i,j}^*} = 0, i = 1, \ldots, m, j = 1, \ldots, n_i.$$

The full MDSA algorithm is described in Algorithm 1, which possesses the following convergence guarantee.

Theorem 5 Suppose there exists a unique optimal solution $(\mathbf{w}_1^*, \ldots, \mathbf{w}_m^*)$ for (5) such that for any $(\mathbf{w}_1, \ldots, \mathbf{w}_m)$ in the feasible set

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i} \frac{\partial Z}{\partial w_{i,j}}(\mathbf{w}_1, \ldots, \mathbf{w}_m)(w_{i,j} - w_{i,j}^*) = 0 \text{ if and only if } \mathbf{w}_i = \mathbf{w}_i^* \text{ for all } i. \tag{14}$$

Then the sequence $\left(\mathbf{w}_1^k, \ldots, \mathbf{w}_m^k\right)$ generated by Algorithm 1 with step size γ_k such that

$$\sum_{k=1}^{\infty} \gamma_k = \infty, \sum_{k=1}^{\infty} \gamma_k^2 < \infty,$$

converges to $(\mathbf{w}_1^*, \ldots, \mathbf{w}_m^*)$ almost surely.

It is easy to verify that the condition (14) must hold if the objective is convex, provided the optimal solution is unique. Besides convexity, condition (14) also applies to objectives that along any ray starting from the optimal solution is a strictly increasing function.

4 NUMERICAL EXPERIMENT

In this section we present some simulation study on the validity of the proposed method, and compare our method with bootstrap resampling (Barton and Schruben 1993, Barton and Schruben 2001). We will show that our method produces statistically valid CIs that have similar coverage, and in some sense are more stable compared with the bootstrap.

We consider the canonical M/M/1 queue with arrival rate 0.8 and service rate 1. The system is empty when the first customer comes in. We are interested in the probability that the 20-th customer waits for

Algorithm 1 MDSA for solving (5)

Input: a parameter $\eta > 0$, an initial feasible solution $(\mathbf{w}_1^1, \ldots, \mathbf{w}_m^1)$, a step size sequence γ_k, and number of replications per iteration R.

Iteration: set $k = 1$

1. Estimate $\hat{\psi}_{i,j}^k = \hat{\psi}_{i,j}(\mathbf{w}_1^k, \ldots, \mathbf{w}_m^k), i = 1, \ldots, m, j = 1, \ldots, n_i$ using

$$\hat{\psi}_{i,j}^k = \frac{1}{R} \sum_{r=1}^{R} h(\mathbf{X}_1^r, \ldots, \mathbf{X}_m^r) S_{i,j}(\mathbf{X}_i^r)$$

where $\mathbf{X}_1^r, \ldots, \mathbf{X}_m^r, r = 1, \ldots, R$ are R independent replications of the m input processes under distributions $(\mathbf{w}_1^k, \ldots, \mathbf{w}_m^k)$

2. Compute the optimal solution $(\lambda_1^{k+1}, \ldots, \lambda_m^{k+1}, \beta^{k+1})$ to

$$\begin{array}{ll} \max & g(\boldsymbol{\lambda}, \beta) \\ \text{subject to} & \boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m) \in \mathbb{R}^m, \beta \geq 0 \end{array}$$

as discussed in Section 3.2, with $\xi_{i,j} = \gamma_k \hat{\psi}_{i,j}^k, u_{i,j} = w_{i,j}^k$.

3. Solve the equations

$$\gamma_k \hat{\psi}_{i,j}^k + \lambda_i^{k+1} + \log \frac{w_{i,j}^{k+1}}{w_{i,j}^k} + 1 - \frac{2\beta^{k+1}}{w_{i,j}^{k+1}} = 0$$

for $(\mathbf{w}_1^{k+1}, \ldots, \mathbf{w}_m^{k+1})$, then go back to 1.

longer than 2 units of time to get served. To put it in the form of (1), let A_t be the inter-arrival time between the t-th and $(t+1)$-th customers, S_t be the service time of the t-th customer, and

$$h(A_1, A_2, \ldots, A_{19}, S_1, S_2, \ldots, S_{19}) = \mathbf{1}_{\{W_{20} > 2\}}, \tag{15}$$

where the waiting time W_{20} is calculated via the Lindley recursion

$$W_1 = 0, W_{t+1} = \max\{W_t + S_t - A_t, 0\}, \text{ for } t = 1, \ldots, 19. \tag{16}$$

To test the method, we pretend that both the inter-arrival time distribution and service time distribution are unknown, but data for both inputs are accessible. Specifically, we generate two samples of size n_1 and n_2 from exponential distribution with rate 0.8 and 1 respectively, and plug in the data into (2) to compute a 95% CI using Algorithm 1.

The parameters of Algorithm 1 are empirically chosen based on several test runs. In all the following experiments, the step size is set to be $\gamma_k = 1/(2\sqrt{n_1 k})$, and the number of replications for gradient estimation $R = 30$. The algorithm is terminated if the difference measured in 1-norm between the average of the last 50 iterates and that of the last 51 to 100 iterates is less than some small $\varepsilon > 0$, which shall be specified in each of the experiments below. The average of the last 50 iterates also serves as the final output of the algorithm.

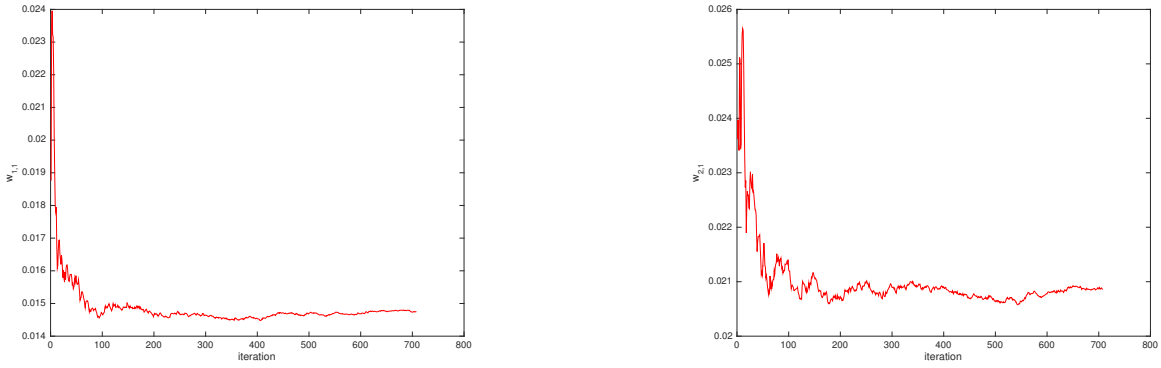
Apart from the stochastic nature of the algorithm, another source of uncertainty that affects the CI is the final evaluation of the objective at the confidence bounds, i.e. computing $\mathscr{L}_\alpha, \mathscr{U}_\alpha$. The bounds are computed by taking the average of a number of replications, called R_e, of (15). Since the length of the CI with only input uncertainty is of order $1/\sqrt{n_1}$, to make the stochastic error negligible, R_e is chosen to be moderately larger than n_1, n_2. The value of R_e for each experiments is also to be specified below.

4.1 Accuracy of the EL confidence interval

We draw 100 samples of sizes $n_1 = n_2 = 30, n_1 = n_2 = 50, n_1 = n_2 = 100$, respectively, and construct one CI based on each sample. Table 1 shows that the CIs have accurate coverage probabilities among all the considered sample sizes, and can be computed in a reasonably short time. Here $R_e = 1000$ is large enough because the sample size is no more than 100. To verify that the algorithm has indeed converged under these stopping criteria ε, trace plots of components of the weight vector are extracted. See Figure 1 for some of them.

Table 1: Performance of EL method under various sample sizes. # of replications for final evaluation $R_e = 1000$ for all, stopping criterion parameter ε are specified in each case.

Sample size	$30, \varepsilon = 0.0073$	$50, \varepsilon = 0.0057$	$100, \varepsilon = 0.0040$
Run time per CI(seconds)	15.1	18.1	22.7
Coverage probability estimate	0.94	0.92	0.94
95% CI for coverage probability	$[0.893, 0.987]$	$[0.867, 0.973]$	$[0.893, 0.987]$
# of replications of h per CI	2.7×10^4	3.4×10^4	4.5×10^4
# of iterations per CI(min+max)	837	1058	1430



(a) $w_{1,1}$ (b) $w_{2,1}$

Figure 1: Trace plot of the first component of the weight vectors for inter-arrival time and service time.

4.2 Comparison with bootstrap resampling

The bootstrap has been widely used to approximate sampling distribution of a statistic from which CIs can be constructed. One method of constructing CIs using the bootstrap is the percentile bootstrap used in Barton and Schruben (1993) and Barton and Schruben (2001). Given a pair of samples, $A^1, A^2, \ldots, A^{n_1}$ for the inter-arrival time and $S^1, S^2, \ldots, S^{n_2}$ for the service time, the percentile bootstrap proceeds as follows. First choose B, the number of bootstrap samples and N, the number of simulation replications for each bootstrap sample. Then for each $b = 1, 2, \ldots, B$, (1) draw a simple random sample of size n_1 with replacement from $\{A^1, \ldots, A^{n_1}\}$ and a sample of size n_2 from $\{S^1, \ldots, S^{n_2}\}$, denoted by $\{A_b^1, \ldots, A_b^{n_1}\}$, and $\{S_b^1, \ldots, S_b^{n_2}\}$ respectively; (2) generate N replications of h with the inter-arrival distribution being uniform over $A_b^1, \ldots, A_b^{n_1}$ and the service time distribution being uniform over $S_b^1, \ldots, S_b^{n_2}$, and take the average Z_b as the estimate of $\mathbb{E}h$. Finally output the $0.025(B+1)$-th and $0.975(B+1)$-th order statistics of $\{Z_b\}_{b=1}^B$, $Z_{(\lfloor 0.025(B+1) \rfloor)}$ and $Z_{(\lfloor 0.975(B+1) \rfloor)}$, as the lower and upper limits of the CI.

To compare our method with bootstrap resampling, we study two cases. In the first case (Table 2), we draw 100 samples of size $n_1 = n_2 = 50$, and compute CIs for each of them. In the second case (Table 3), we draw only one sample of size $n_1 = n_2 = 50$, and repeat computing 50 CIs based on a *single* sample. To make the comparison fair, we appropriately set the parameters B, N for the bootstrap method, and ε, R_e for the EL method so that the run times are comparable. The result shows that the EL method generates comparable but slightly narrower CIs than the bootstrap, while still keeps similar coverage probabilities. In the first case, most of the uncertainty in either method comes from the input data, so the CIs they generate exhibit similar variation. In the second case, the input data is fixed, and hence the fluctuations of the CIs are solely due to the resampling and the stochastic noises. Table 3 shows that the EL method gives more stable CIs than the bootstrap, meaning that the EL is less vulnerable to these noises. This can be attributed to the fact that EL is an optimization-based method and hence does not succumb to the additional resampling noise introduced in the bootstrap.

Table 2: EL method versus bootstrap, 100 confidence intervals for 100 samples.

	Bootstrap		EL method
	$B = 1000, N = 500$	$B = 500, N = 1000$	$\varepsilon = 0.0057, R_e = 500$
Run time per CI(seconds)	17.4	17.4	18.1
Coverage probability estimate	0.90	0.94	0.93
95% CI for coverage probability	$[0.841, 0.959]$	$[0.893, 0.987]$	$[0.880, 0.980]$
Mean CI length	0.595	0.606	0.536
Standard deviation of CI length	0.094	0.086	0.080
Mean lower limit	0.166	0.155	0.196
Standard deviation of lower limit	0.121	0.093	0.105
Mean upper limit	0.761	0.760	0.733
Standard deviation of upper limit	0.157	0.142	0.133
# replications of h per CI	5×10^5	5×10^5	3.3×10^4

Table 3: EL method versus bootstrap, 50 replications of confidence intervals for a single sample.

	Bootstrap,$B = 500, N = 1000$	EL,$\varepsilon = 0.0064, R_e = 2 \times 10^5$
Run time per CI(seconds)	17.2	18.9
Standard deviation of CI length	0.0233	0.0053
Standard deviation of lower limit	0.0106	0.0038
Standard deviation of upper limit	0.0213	0.0030
# replications of h per CI	5×10^5	4.3×10^5

5 CONCLUSION

We have proposed an optimization-based method to quantify nonparametric input uncertainty for simulation performance measures. The method computes CIs that account for input errors by positing a pair of optimizations subject to a weighted average of Burg-entropy divergence constraints on the collection of empirical input models. We have argued the statistical accuracy of these optimizations via the EL method. We have also investigated an MDSA algorithm to locally solve the optimizations, and compared its numerical performances with bootstrap resampling. In future work, we will investigate the generalization of this approach to other types of performance measures, and more efficient algorithms to solve the involved class of optimizations.

ACKNOWLEDGMENTS

We gratefully acknowledge support from the National Science Foundation under grants CMMI-1400391/1542020 and CMMI-1436247/1523453.

REFERENCES

Barton, R. R., and L. W. Schruben. 1993. "Uniform and Bootstrap Resampling of Empirical Distributions". In *Proceedings of the 1993 Winter Simulation Conference*, edited by G. Evans, M. Mollaghasemi, E. Russell, and W. Biles, 503–508. ACM.

Barton, R. R., and L. W. Schruben. 2001. "Resampling Methods for Input Modeling". In *Proceedings of the 2001 Winter Simulation Conference*, edited by B. A. Peters, J. S. Smith, D. J. Medeiros, and M. W. Rohrer, Volume 1, 372–378. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Beck, A., and M. Teboulle. 2003. "Mirror Descent and Nonlinear Projected Subgradient Methods for Convex Optimization". *Operations Research Letters* 31 (3): 167–175.

Ben-Tal, A., D. Den Hertog, A. De Waegenaere, B. Melenberg, and G. Rennen. 2013. "Robust Solutions of Optimization Problems Affected by Uncertain Probabilities". *Management Science* 59 (2): 341–357.

Cox, D. R., and D. V. Hinkley. 1979. *Theoretical statistics*. CRC Press.

Delage, E., and Y. Ye. 2010. "Distributionally Robust Optimization Under Moment Uncertainty with Application to Data-Driven Problems". *Operations Research* 58 (3): 595–612.

Ghosh, S., and H. Lam. 2015a. "Computing Worst-Case Input Models in Stochastic Simulation". *arXiv preprint arXiv:1507.05609*.

Ghosh, S., and H. Lam. 2015b. "Mirror Descent Stochastic Approximation for Computing Worst-Case Stochastic Input Models". In *Proceedings of the 2015 Winter Simulation Conference*, edited by S. Jain, R. R. Creasey, J. Himmelspach, K. P. White, and M. Fu, 425–436. Piscataway, New Jersey: IEEE Press: Institute of Electrical and Electronics Engineers, Inc.

Goh, J., and M. Sim. 2010. "Distributionally Robust Optimization and its Tractable Approximations". *Operations Research* 58 (4-part-1): 902–917.

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.

Nemirovsky, A.-S., D.-B. Yudin, and E.-R. Dawson. 1982. *Problem Complexity and Method Efficiency in Optimization.* John Wiley & Sons, Inc., Panstwowe Wydawnictwo Naukowe (PWN).

Owen, A. B. 1988. "Empirical Likelihood Ratio Confidence Intervals for a Single Functional". *Biometrika* 75 (2): 237–249.

Owen, A. B. 2001. *Empirical Likelihood*. CRC Press.

Pardo, L. 2005. *Statistical Inference Based on Divergence Measures*. CRC Press.

AUTHOR BIOGRAPHIES

HENRY LAM is an Assistant Professor in the Department of Industrial and Operations Engineering at the University of Michigan, Ann Arbor. His research focuses on stochastic simulation, risk analysis, and simulation optimization. His email address is khlam@umich.edu.

HUAJIE QIAN is a Ph.D. student in Applied and Interdisciplinary Mathematics at the University of Michigan, Ann Arbor. His research interest lies in applied probability and simulation optimization. His email address is hqian@umich.edu.