Proceedings of the 2015 Winter Simulation Conference
L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, eds.

STOCHASTIC OPTIMIZATION USING HELLINGER DISTANCE

Anand N. Vidyashankar

Department of Statistics
George Mason University
4400 University Drive
Fairfax, VA 22030, USA

Jie Xu

Department of Systems Engineering and Operations Research
George Mason University
4400 University Drive
Fairfax, VA 22030, USA

ABSTRACT

Stochastic optimization facilitates decision making in uncertain environments. In typical problems, probability distributions are fit to historical data for the chance variables and then optimization is carried out, as if the estimated probability distributions are the "truth". However, this perspective is optimistic in nature and can frequently lead to sub-optimal or infeasible results because the distribution can be misspecified and the historical data set may be contaminated. In this paper, we propose to integrate existing approaches to decision making under uncertainty with robust and efficient estimation procedures using Hellinger distance. Within the existing decision-making methodologies that make use of parametric models, our approach offers robustness against model misspecifications and data contamination. Additionally, it also facilitates quantification of the impact of uncertainty in historical data on optimization results.

1 INTRODUCTION

Stochastic optimization and its variants have received much attention during the last few decades. An important aspect of the stochastic optimization problem is that it facilitates decision making in uncertain environments. In typical problems, historical data are used to estimate the probability distributions for the random variables and then optimization is carried out, assuming that the estimated probability distributions are the "nominal". However, this perspective is optimistic in nature and can frequently lead to sub-optimal or infeasible results because the distribution may be misspecified and parameter estimates may have significant uncertainty due to the limited size of the data set and/or data contamination. This has been adequately described in the literature (Scarf et al. 1958, Ben-Tal and Nemirovski 1998, Ben-Tal and Nemirovski 2000, Bertsimas and Sim 2004). To address this issue these authors introduced the concept of robust optimization (RO). In this approach the goal is to find an optimal solution, for the parameters governing the optimization problem, which are immune to ambiguity in the parameters. The method models the ambiguity by restricting the parameters to a set, referred to as an uncertainty set and the optimization is carried out under the worst case scenario.

An alternative approach, referred to as distributionally robust optimization (DRO), accounts for stochastic nature of the parameters. In this method, a stochastic optimization problem is considered, where the

distributions of the parameters are allowed to vary in an ambiguity set. The optimization is then carried out under the worst case distribution in the ambiguity set. An important question concerns the construction of ambiguity/uncertainty sets (we use "ambiguity set" and "uncertainty set" interchangeably in this paper). One of the earlier methods studied by Scarf et al. (1958) and later extended to more complex objective functions (Yue et al. 2006, Zhu et al. 2006, Popescu 2007), assumed that the first two moments were known. Delage and Ye (2010) studied stochastic optimization allowing for moment uncertainty. They constructed confidence intervals for the mean and the covariance matrix using concentration inequalities of McDiarmid (1998). Additionally, they also studied the usefulness of including support constraints. Recently, Hu and Hong (2013) studied DRO, where the ambiguity set was determined by Kullback-Leibler divergence between the probability distributions and the "nominal" distribution. It is worth pointing out here that the nominal distribution is based on historical data and hence is typically estimated.

The impact of input uncertainty has also been recently studied in the stochastic simulation literature under a likelihood scenario (Cheng and Holland 1997, Chick 2001, Zouaoui and Wilson 2003, Zouaoui and Wilson 2004, Ng and Chick 2006, Barton et al. 2014, Xie et al. 2013, Xie et al. 2014). The focus is on quantifying the impact of parametric input distribution model uncertainty, arising from historical data uncertainty, on simulation estimations in the presence of simulation noise in output. Also related is a rich body of literature on measuring parametric uncertainty via derivative estimation (L'Ecuyer 1990, Ho and Cao 1991, Glasserman 1991, Fu and Hu 1997, Fu 1994, Glasserman and Tayur 1995, Hong 2009, Hong and Liu 2009). Alternatively, Lam (2013a), Lam (2013b) consider a non-parametric approach in which one evaluates the sensitivity in a non-parametric "neighborhood" around the "true" distribution. This neighborhood is defined using Kullback-Leibler (KL) divergence between the probability distributions. All these methods provide useful assessment of distribution uncertainty for a *fixed* decision variable, but do not consider the impact on the *optimizers* and the values of the objective function at the optimizers.

The primary objective of this article is to provide an alternative paradigm for stochastic optimization that provides robustness. Our proposed method does not increase the complexity of the computation. Indeed, in some examples, the method reduces the computational complexity by a significant factor. Our method is based on tools that incorporate statistical robustness and sample average approximation (SAA).

In the rest of the article, we describe the problem in Section 2 and provide two motivating examples. Section 3 provides precise mathematical formulation. Section 4 studies the asymptotic properties of the estimates. Section 5 provides numerical experiment results. Section 6 contains concluding remarks.

2 PROBLEM DESCRIPTIONS

We assume that the decision space $\mathscr{X} \subseteq \mathbb{Z}^d$ where $\mathscr{X}$ is compact and $\mathbb{Z}^d$ is the d-dimensional integer lattice. Let $\xi : \Omega \to \mathbb{R}$ be a random variable. Let $h(\mathbf{x}, \xi) : \mathscr{X} \times \Omega \to \mathbb{R}$ be a real-valued Borel measurable function. Let $Y_1, Y_2, \ldots, Y_n$ denote the historical data set, which consists of a collection of n independent and identically distributed (i.i.d.) random variables with distribution μ_G. Assume that $g(\cdot)$ is the density of μ_G. Let the postulated distribution of Y_1 be μ_{θ_0} with density $f(y;\theta_0)$ belonging to the family $\mathscr{F} = \{f(y;\theta); \theta \in \Theta\}$. The problem we study concerns estimating

$$\max_{\mathbf{x}\in\mathscr{X}} \mathbf{E}_\nu \left[h(\mathbf{x};\xi)\right] \stackrel{\Delta}{=} M, \tag{1}$$

and

$$\arg\max_{\mathbf{x}\in\mathscr{X}} \mathbf{E}_\nu \left[h(\mathbf{x};\xi)\right] \stackrel{\Delta}{=} \mathbf{x}^*, \tag{2}$$

where ν represents either μ_G or μ_{θ_0}, using $Y_1, Y_2, \ldots, Y_n$. When there are multiple optimizers, we denote the set of solutions of (2) by S^*.

2.1 Motivating Examples.

We consider two examples in this section to motivate our work. The first example is the well-known news-vendor problem in inventory theory (Hopp and Spearman 2011). The second example is a three-stage flowline system, which showcases a queueing system that may be part of a large-scale and complex manufacturing or service system (Buzacott and Shantikumar 1993).

2.1.1 The News Vendor Problem.

In the news-vendor problem, the decision maker needs to decide how many units of a product to order before the selling period begins. There is no replenishment of inventory during the selling period. Let c_o be the cost per unit of product left over after the sale is over, and c_s be the per unit cost of shortage. Let $F(\cdot;\theta)$ be the cumulative distribution function (CDF) for the demand. Then the optimal order quantity is given by

$$\mathbf{x}^* = F^{-1}\left(1 - c_o/(c_s + c_o);\theta\right), \tag{3}$$

In practice, the decision maker often postulates a distribution model $F(\cdot;\theta)$ and estimates θ from a data set (historical sale records) of size n, denoted as $\widehat{\theta}_n$. Equation (3) shows that the optimal order quantity $\mathbf{x}^*$ is a function of the postulated CDF $F(\cdot;\theta)$ and the parameter estimate $\widehat{\theta}_n$, which is a function of the data set. We make this dependence explicit by writing $\widehat{\mathbf{x}}_n^*$ in (3), where n is the number of samples in the data set.

2.1.2 Flowline Optimization.

In this example, we consider optimizing the design of an N-stage flowline (Buzacott and Shantikumar 1993, Pichitlamken and Nelson 2003), as illustrated in Figure 1. There are finite buffer storage spaces in front of stations $2,\ldots,N$, denoted by $b_2,\ldots,b_N$. The total number of buffer spaces $\sum_{i=1}^{N} b_i$ cannot exceed B. There is an infinite number of jobs in front of station 1 for processing. We adopt the production blocking policy (Buzacott and Shantikumar 1993), i.e., each station will serve a job as long as there is a job available and the station is not blocked, that is, the job it has completed cannot be released to the downstream station because the buffer for that station is full.

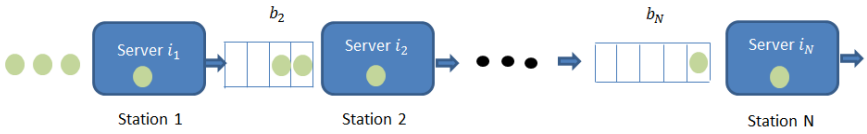

Figure 1: A N-stage flowline system.

Each station has a single server and there are N flexible servers that can be allocated to each of these N stations. The service time distributions of these N servers have CDFs $F_1(\cdot;\theta_1), F_2(\cdot;\theta_2),\ldots,F_N(\cdot;\theta_N)$. The objective is to allocate these N servers to N stations and determine the number of buffer spaces in front of stations 2 to N such that the steady-state throughput is maximized.

In the special case of $N = 3$, if service time distributions can be ordered by likelihood ratio, the optimal server allocation calls for the allocation of the fastest server to station 2 (Buzacott and Shantikumar 1993). The allocations of the other two servers do not matter. With this server allocation decision, one can then solve the optimal buffer space allocation b_2^* and b_3^*, subject to the total buffer space constraint $b_2^* + b_3^* \le B$. When the service time distributions are postulated to be exponential with mean service times $\mu_i, i = 1,2,3$, we can analytically calculate the throughput of the flowline for a buffer allocation by solving the Markov state balance equations for the flowline, parameterized by $\mu_i, i = 1,2,3$ (Buzacott and Shantikumar 1993). One can then identify the allocation with the maximum throughput.

In reality, one has to estimate $\mu_i, i = 1,2,3$ from data sets recording the service times for each of the three servers. Therefore, the identity of the fastest server is not certain and really depends on the data set. Since Markov state balance equations are parameterized by the estimates $\widehat{\mu}_i, i = 1,2,3$, the optimal buffer sizes are also a function of data set, and we make this dependence on data set explicit by denoting them as $\widehat{b}^*_{2,n}$ and $\widehat{b}^*_{3,n}$.

When the service time distributions can still be ordered by likelihood ratio but are not exponential, e.g., Pareto random variables, we can no longer analytically determine the throughput and thus have to use an optimization via simulation approach to find the optimal buffer space allocation (Xu et al. 2010).

3 ESTIMATORS AND ESTIMATED OPTIMIZERS

Recall that $\mathscr{Y}_n = (Y_1, Y_2, \ldots, Y_n)$ denote a collection of n i.i.d. random variables with distribution μ_G and $g(\cdot)$ is the density of μ_G, and is postulated to be of a parametric family $\{f(y;\theta); \theta \in \Theta\}$, where $\Theta \subseteq \mathbb{R}^q$. We first provide a concise description of minimum Hellinger distance estimator (MHDE) in this section, in comparison with the widely used maximal likelihood estimator (MLE).

3.1 Methods of Estimation.

Let $\widehat{\theta}_n$ denote the estimators of θ using the historical data $\mathscr{Y}_n$ which is assumed to have a density $f(\cdot;\theta_0), \theta_0 \in \Theta$. Then we set

$$\begin{aligned} \widehat{\mathbf{x}}^*_n &= \arg\max_{\mathbf{x}\in\mathscr{X}} \mathrm{E}_{\mu_{\widehat{\theta}_n}}[h(\mathbf{x};Y)] \\ &= \arg\max_{\mathbf{x}\in\mathscr{X}} \int_{\mathbb{R}^q} h(\mathbf{x};Y) f(y;\widehat{\theta}_n) dy. \end{aligned}$$

It is common to use for $\widehat{\theta}_n$, the MLE of θ_0 obtained by maximizing

$$L_n(\theta | Y_1, Y_2, \ldots, Y_n) = \prod_{i=1}^n f(Y_i;\theta). \tag{4}$$

In such a situation, it is known that under moment and regularity conditions (Lehmann and Casella 1998)

$$\widehat{\theta}_n \xrightarrow{a.s.} \theta_0,$$

and

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) \xrightarrow{d} N(0, I^{-1}(\theta_0)),$$

where

$$I(\theta_0) = -\mathrm{E}_\theta \left[\frac{\partial^2 \log f(Y;\theta)}{\partial \theta^2}\right] |_{\theta=\theta_0}$$

is the Fisher information matrix. Hence, under the assumption that $h^*(y) = \max_{\mathbf{x}\in\mathscr{X}} |h(\mathbf{x};y)|$ is a bounded function, it follows that under additional smoothness conditions

$$\lim_{n\to\infty} \int h(\mathbf{x};y) f(y;\widehat{\theta}_n) dy = \int h(\mathbf{x};y) f(y;\theta_0) dy, \text{ w. p. 1.}$$

This would imply, due to the finiteness of the decision space $\mathscr{X}$, that

$$\widehat{\mathbf{x}}^*_n = \arg\max_{\mathbf{x}\in\mathscr{X}} \int h(\mathbf{x};y) f(y;\widehat{\theta}_n) dy$$

converges almost surely to $\mathbf{x}^*$. However, if the model is misspecified, the resulting estimator of $\mathbf{x}^*$ will incur a substantial bias as illustrated in Section 5. It is known that MHDE yields efficiency when the model

is correctly specified while is robust to model misspecification and data contamination (Beran 1977, Beran 1978). The Hellinger distance between two densities $f(\cdot)$ and $g(\cdot)$ is given by

$$HD(f,g) = \left(\int_{\mathbb{R}} (\sqrt{f(y)} - \sqrt{g(y)})^2 dy \right)^{1/2}.$$

Then $HD^2(f(\cdot;\theta),g)$ can be used as an objective function that one can minimize with respect to θ; set

$$\theta_g = \arg\min_{\theta\in\Theta} HD^2(f(\cdot;\theta),g).$$

Here θ_g is the population version of MHDE of θ.

Let $g_n(\cdot)$ denote a nonparametric estimator of $g(\cdot)$. A useful choice for $g_n(\cdot)$ is the kernel density estimators, namely

$$g_n(y) = \frac{1}{nc_n} \sum_{i=1}^{n} K\left(\frac{y - Y_i}{c_n} \right),$$

where $K(\cdot)$ is a kernel function (e.g., Gaussian kernel) and c_n is referred to as the bandwidth or window width and $c_n \to 0$, $nc_n \to \infty$ as $n \to \infty$. Then the sample version of the MHDE is

$$\tilde{\theta}_n = \arg\min_{\theta\in\Theta} HD^2(f(\cdot;\theta),g_n).$$

The above minimization problem is equivalent to the following maximization problem; namely,

$$\tilde{\theta}_n = \arg\max_{\theta\in\Theta} \int_{\mathbb{R}} f^{1/2}(y;\theta) g_n^{1/2}(y) dy. \tag{5}$$

Then, it is well-known that $\tilde{\theta}_n$ converges w.p. 1 to θ_g. When the postulated model coincides with the true model, i.e., $f(\cdot;\theta_0) \equiv g$, then $\theta_g = \theta_0$. Additionally, under additional regularity conditions, it can be shown

$$\sqrt{n}\left(\tilde{\theta}_n - \theta_g\right) \xrightarrow{d} N\left(0, \frac{1}{4}\int \rho_g(y)\rho_g(y)^T dy\right), \tag{6}$$

where

$$\rho_g(y) = -\left[\int \ddot{s}_\theta(y)\sqrt{g(y)}dy\right]^{-1} \left[\nabla_\theta f(y;\theta)|_{\theta=\theta_g}\right],$$

where $s_\theta(y) = \sqrt{f(y;\theta)}$ and $\ddot{s}_\theta(\cdot)$ is the matrix of second partial derivatives. For notational simplicity, we use V to denote the covariance matrix of the multivariate normal distribution in (6). When $g = f(\cdot;\theta_0)$, then $V = I^{-1}(\theta_0)$, where $I(\theta_0)$ is the Fisher information matrix (Beran 1977, Cheng and Vidyashankar 2006, Sriram and Vidyashankar 2000, Chan 2008).

It is helpful to compare the estimation equations (4) and (5) to understand why MHDE is robust. We approximate the integrals in (4) and (5) by summations at $Y_1, Y_2, \ldots, Y_n$. For MHDE, this leads to $\tilde{\theta}_n \approx \arg\max_{\theta\in\Theta} \sum_{i=1}^n f^{1/2}(Y_i;\theta) g_n^{1/2}(Y_i)$. The first order condition requires solving the equation

$$\sum_{i=1}^{n} u(Y_i;\theta) f^{1/2}(Y_i;\theta) g_n^{1/2}(Y_i) = 0, \tag{7}$$

where $u(Y_i;\theta) = \dot{f}(Y_i;\theta)/f^{1/2}(Y_i;\theta)$ is the score function, and $g_n(Y) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}_{\{Y=Y_i\}}$. In comparison, MLE solves

$$\sum_{i=1}^{n} u(Y_i;\theta) g_n(Y_i) = 0. \tag{8}$$

Comparing (7) and (8), we see MHDE weighs each observation Y_i by the postulated density $f^{1/2}(Y_i;\theta)$, while MLE assigns equal weight to all observations. When MHDE sees outliers, it will apply smaller weights to these outliers and dampen their influence, and thus achieves robustness.

3.2 Estimation of the Optimizer and Optimal Objective Value.

For a fixed $\mathbf{x}$, an estimator of $\mathrm{E}_{\mu_{\theta_g}}(h(\mathbf{x},Y))$ is given by

$$\mathrm{E}_{\mu_{\tilde{\theta}_n}}(h(\mathbf{x},Y)) = \int_{\mathbb{R}} h((\mathbf{x},y))f(y|\tilde{\theta}_n)dy. \tag{9}$$

In general, the integral in (9) is difficult to evaluate. One can then adopt SAA (Kleywegt et al. 2002, Nemirovski et al. 2009) to obtain the approximation

$$A_n(\mathbf{x}) \equiv \frac{1}{m_n}\sum_{i=1}^{m_n} h(\mathbf{x},Y_{n,i}), \tag{10}$$

where $Y_{n,i}$ is the ith i.i.d. sample drawn from the fitted parametric model $f(y|\tilde{\theta}_n)$.

Recall that the set of solutions to (2) is denoted by S^*. Now let

$$S^*_{n,m_n} = \{\mathbf{x}|A_n(\mathbf{x}) \text{ is maximized}\}.$$

Also let S^*_ε and $S^*_\varepsilon(n,m_n)$ denote the ε-optimal solutions to (2) and (11) respectively, i.e.,

$$S^*_\varepsilon = \{\mathbf{x}\in\mathbf{X}|\mathrm{E}_{\mu_{\theta_0}}(h(\mathbf{x},Y)) \le \mathrm{E}_{\mu_{\theta_0}}(h(\mathbf{x}',Y)) + \varepsilon, \forall \mathbf{x}'\in\mathbf{X}, \mathbf{x}'\neq\mathbf{x}\},$$

and

$$S^*_\varepsilon(n,m_n) = \{\mathbf{x}\in\mathbf{X}|A_{n,m_n}(\mathbf{x}) \le A_{n,m_n}(\mathbf{x}') + \varepsilon, \forall \mathbf{x}'\in\mathbf{X}, \mathbf{x}'\neq\mathbf{x}\}.$$

4 PROPERTIES OF ESTIMATORS AND ESTIMATED OPTIMIZERS

Our first result is concerned with the uniform approximation of (9) by (10).

Theorem 1 Assume that (i) the decision space $\mathbf{X}$ is a finite discrete space; (ii) $\max_{\mathbf{x}\in\mathbf{X}}\mathrm{E}_{\mu_g}(|h(\mathbf{x},Y)|) < \infty$; (iii) the density function of the historical data satisfies the regularity conditions of Vidyashankar and Xu (2015); and (iv) $c_n \to 0, nc_n \to \infty$ hold. Then

$$\sup_{\mathbf{x}\in\mathbf{X}} |A_n(\mathbf{x}) - \mathrm{E}_{P_{\tilde{\theta}_n}}(h(\mathbf{x},Y))| \xrightarrow{a.s.} 0 \text{ as } n\to\infty \text{ and } m_n\to\infty,$$

$$\sup_{\mathbf{x}\in\mathbf{X}} |A_n(\mathbf{x}) - \mathrm{E}_{P_{\theta_g}}(h(\mathbf{x},Y))| \xrightarrow{a.s.} 0 \text{ as } n\to\infty \text{ and } m_n\to\infty.$$

We notice here that the result improves and sharpens the work of Kleywegt et al. (2002) by explicitly taking into account data-driven aspects. Traditionally in SAA, Monte Carlo approximations to the integral using a *fixed* probability distribution is performed and then the optimization is carried out on the approximated objective function. However, in our approach, the approximation changes with n, the size of the historical data. Hence the behavior of the optimizer needs to be evaluated over the distribution of the historical data. This brings into play the role of (i) sample size n and (ii) the estimate of the density, both of which are addressed in Theorem 1. Our next result is concerned with the behavior of the objective functions.

Theorem 2 The following hold:

(i) $\lim_{n\to\infty}\min_{\mathbf{x}\in\mathbf{X}} A_n(\mathbf{x}) = \min_{\mathbf{x}\in\mathbf{X}}\mathrm{E}_{\theta_g}(h(\mathbf{x},Y))$ with probability 1;

(ii) For any $\varepsilon > 0$, there exists a N_ε (which is typically random) such that $\forall n \ge N_\varepsilon$, $S^*_\varepsilon(n,m_n) \subset S^*_\varepsilon$ with probability 1.

We now turn our attention to the properties of the minimizer. To provide a succinct description of the problem, we need to introduce some aspects of large deviation theory (Dembo and Zeitouni 1998). To this end, we study the random function

$$H_D(\mathbf{x},Y) = h(\mathbf{x}^*,Y) - h(\mathbf{x},Y),$$

where $\mathbf{x}^* = \arg\max \mathrm{E}_{\mu_{\theta_g}}[h(\mathbf{x},Y)]$, and we assume here that S^* is a singleton. Under exponential moment hypothesis and steepness of $K_{\mathbf{x}}(\theta) = \log \mathrm{E}[e^{\theta H_D(\mathbf{x})}]$ (Dembo and Zeitouni 1998), one can establish a large-deviation principle for the quantity

$$T_n(\mathbf{x}) \equiv \frac{1}{m_n}\sum_{i=1}^{m_n} H_D(\mathbf{x},Y_{n,i}).$$

Indeed, the main result is that $\{T_n(\mathbf{x}), n \geq 1\}$ satisfies the large-deviation principle with rate function

$$\Lambda^*_{\mathbf{x}}(u) = \sup_{\theta}[\theta u - K_{\mathbf{x}}(\theta)].$$

A significant new difficulty arises here due to the triangular array nature of the random variables $Y_{n,i}$ and the role of the kernel density estimates. Our main result concerning the minimizers is that

Theorem 3 Under the assumption that $H_D(\mathbf{x},Y)$ possesses a steep generating function, the minimizer $\widehat{\mathbf{x}}^*_n = \arg\min_{\mathbf{x}\in\mathbf{X}} \frac{1}{m_n}\sum_{i=1}^{m_n} h(\mathbf{x},Y_{n,i})$ converges to $\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathbf{X}} \mathrm{E}_{\mu_{\theta_g}}(h(\mathbf{x},Y))$, and that

$$\lim_{n\to\infty}\frac{1}{n}\log(1-P(\widehat{\mathbf{x}}^*_n = \mathbf{x}^*)) = -\inf_{\mathbf{x}\in\mathbf{X},\mathbf{x}\neq\mathbf{x}^*}\Lambda^*_{\mathbf{x}}(0).$$

The significance of the above theorem is that it helps identify the historical data sample size so that the probability that the optimizer corresponding to the estimated objective function is different from the true optimizer is small. Once again we notice that the probability distribution is based on the historical data and not a fixed probability distribution as is typically done in stochastic optimization. We now turn our attention to objective values. Let $\sigma^2(\mathbf{x}) = \mathrm{Var}(h(\mathbf{x},Y))$.

Theorem 4 Assume that $\sigma^2(\mathbf{x})$ is finite for all $\mathbf{x}\in\mathbf{X}$. Let $\widehat{\mathbf{x}}^*_n = \arg\min_{\mathbf{x}\in\mathbf{X}} A_n(\mathbf{x})$. Then under regularity conditions in Cheng and Vidyashankar (2006) (see also Hooker and Vidyashankar (2014))

$$\sqrt{n}(\widehat{\mathbf{x}}^*_n - \mathbf{x}^*) \xrightarrow{d} \min_{\mathbf{x}\in S^*} Z(\mathbf{x}),$$

where for every $\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_k \in \mathbf{X}$, $(Z(\mathbf{x}_1),\ldots,Z(\mathbf{x}_k)) \xrightarrow{d} N_k(\mathbf{0},\Sigma_{kk})$, with Σ_{kk} being the covariance matrix of $h(\mathbf{x}_1,Y),\ldots,h(\mathbf{x}_k,Y)$.

It is critical to notice that all our statements are with respect to the distribution of the historical data. A critical issue here is that SAA approximation itself is based on simulations from the fitted parametric model and is reminiscent of the parametric bootstrap method. However, the asymptotic behavior of the parametric bootstrap beyond MLE has remained open for a long time.

5 NUMERICAL EXPERIMENTS

In this section, we report results of applying MHDE to optimize the order quantity in the newsvendor problem described in Section 2.1.1 and the flowline design described in Section 2.1.2. Notice that the decision $\widehat{\mathbf{x}}^*_n$ and the associated objective values $\widehat{M}^*_n \equiv \mathrm{E}_{\mu_{\widehat{\theta}_n}}[h(\widehat{\mathbf{x}}^*_n;\xi)]$ are random variables, depending on the data set and model specification. For each example, we report four types of statistics:

- The sample mean of $\widehat{\mathbf{x}}^*_n$;

- The mean squared error (MSE) of $\widehat{\mathbf{x}}_n^*$;
- The mean *relative optimality gap* (ROG) of $\widehat{M}_n^*$;
- The empirical probability distribution of the ROG of $\widehat{M}_n^*$.

We define ROG as follows. Recall we denote the *true* optimal objective value as M. Let

$$\overline{M} \equiv \frac{1}{|\mathscr{X}|} \sum_{\mathbf{x} \in \mathscr{X}} \mathrm{E}_{\mu_{\widehat{\theta}_n}}[h(\widehat{\mathbf{x}}; \xi)],$$

where $|\mathscr{X}|$ is the number of feasible solutions. Then

$$ROG \equiv \frac{\left|M - \widehat{M}_n\right|}{\left|M - \overline{M}\right|}.$$

In words, $\overline{M}$ represents the expected quality of the decision if one does not perform any optimization and simply randomly pick a feasible solution $\mathbf{x} \in \mathscr{X}$. The numerator in (11) measures the expected improvement of optimization with *perfect* information. The denominator measures how much we lose by using *imperfect* information. It is worthwhile to notice that *ROG* may be larger than 1, indicating that the "optimal" decision based on the postulated model with estimated parameters $\mu_{\widehat{\theta}_n}$ is actually *worse* than randomly picking a feasible decision. We propose to use *ROG* instead of the perhaps more commonly used optimality gap $|M - \widehat{M}_n|/|M|$ because *ROG* is less problem dependent (i.e., the scale of $|M|$) and thus helps better illustrate how much optimization benefit we lose due to imperfect information.

5.1 Newsvendor Problem

We conduct the following three numerical experiments with $c_s = 600, c_o = 400$. For each experiment, we generate $m = 5000$ data sets. We fit a normal distribution to these data sets using MHDE and MLE. The estimated $\tilde{\theta}_n$, and $\widehat{\theta}_n$ are then plugged into (3) to calculate $\widehat{x}_n^*$. For MHDE, Epanechnikov kernel was used with the bandwidth c_n calculated using Silverman's plug-in bandwidth approach $c_n = 2.34 \times \widehat{\sigma} \times n^{-1/5}$, where $\widehat{\sigma}$ is the sample standard deviation of the data set.

- *Baseline*. $D \sim N(30, 9)$. The data set is n i.i.d. realizations of $N(30, 9)$, and $x^* = 30.8$.
- *Model misspecification*. We consider the case where D has a shifted lognormal distribution $28.97 + LN(\mu = -1.096, \sigma^2 = 2.250)$. The data set is n i.i.d. realizations of this distribution. D still has a mean of 30 and a variance of 9. But now the true optimal protection level is $x^* = 29.5$.
- *Data contamination*. $D \sim N(30, 9)$. The data set is generated from a normal mixture $0.9N(30, 9) + 0.1N(60, 9)$. $x^* = 30.8$, since the underlying true distribution is the same as in the baseline case.

The sample means and mean squared errors (MSE) of $\widehat{\mathbf{x}}_n^*$ are reported in Tables 1 and 2. We observe that MHDE estimators are asymptotically efficient as their MSE values are very close to those of MLE. The reason that the MSE values of MHDE are slightly higher is we used a Monte Carlo procedure to obtain MHDE as explained in detail in Vidyashankar and Xu (2015). MHDE make the optimization results robust against both model misspecifications and data contamination. In comparison, the MLE approach suffers from biases caused by model misspecification and data contamination, setting $\widehat{\mathbf{x}}_n^*$ larger than the true optimal values.

To calculate *ROG*, we need to specify the feasible decision space $\mathscr{X}$. We let $\mathscr{X} = [\mu - 5\sigma, \mu + 5\sigma]$, where μ and σ are the mean and standard deviation of the demand. We believe this represents a large and yet fairly reasonable decision space as an inventory manager would hardly have an order quantity beyond $\pm 5\sigma$ of the forecast demand with non-negligible shortage and overstock cost. Table 3 reports the results. From the table, we see that model misspecification and data contamination has substantial

Table 1: Sample Means of $\widehat{\mathbf{x}}_n^*$

	Baseline		Lognormal		Contamination	
n	MHDE	MLE	MHDE	MLE	MHDE	MLE
25	30.8	30.7	29.8	30.5	31.1	36.1
50	30.8	30.7	29.8	30.5	31.0	35.4
100	30.8	30.8	29.7	30.6	30.9	35.4

Table 2: Mean Squared Errors of $\widehat{\mathbf{x}}_n^*$

	Baseline		Lognormal		Contamination	
n	MHDE	MLE	MHDE	MLE	MHDE	MLE
25	.387	.377	.196	2.00	.568	29.2
50	.198	.193	.0963	1.63	.254	21.5
100	.0943	.0925	.0482	1.484	.133	21.5

Table 3: The Sample Means of *ROG* (in %) of the Newsvendor Problem

	Baseline		Lognormal		Contamination	
n	MHDE	MLE	MHDE	MLE	MHDE	MLE
25	0.91	0.91	0.51	3.17	1.28	45.95
50	0.48	0.49	0.18	2.92	0.57	36.28
100	0.24	0.27	0.08	2.86	0.24	35.97

impact on the actual quality of the decisions when MLE estimates are used. This is especially so with the data contamination case, with decisions made based on MLE estimates lose about half of the benefit of optimization.

5.2 Flowline Results

We study the allocation of three servers to the three stations and the allocation of a total of 6 buffer spaces to b_2 and b_3. Sample sizes are set to $n = 25, 50$ and 100. For each experiment, we generate $m = 5000$ data sets and fit exponential distribution to the data sets. We then calculate the estimates of μ_1, μ_2, μ_3 for each data set, $\widehat{\mu}_{1,n}, \widehat{\mu}_{2,n}, \widehat{\mu}_{3,n}$ via MHDE or MLE. Because of the exponential distribution assumption, we would place the machine with the smallest estimated mean service time in the middle station. The optimal buffer sizes $\widehat{b}_{2,n}^*$ and $\widehat{b}_{3,n}^*$ are found by numerically solving the Markov balance equations parameterized by $\widehat{\mu}_{1,n}, \widehat{\mu}_{2,n}, \widehat{\mu}_{3,n}$. We compare optimization results under three settings.

- *Baseline*. Data sets contain i.i.d. observations generated from exponential distributions with mean service times $\mu_1 = 0.5, \mu_2 = 1, \mu_3 = 1.5$. The true optimal flowline design has the server 1 in the middle, server 2 in the first station, and server 3 in the last station. Then we have the optimal buffer allocation as $b_2^* = 2, b_3^* = 3$.
- *Model misspecification*. Service times of all three servers follow Pareto distributions: $Pareto(1.25, 0.1)$, $Pareto(1.25, 0.2)$, and $Pareto(1.25, 0.3)$. The mean service times are the same as in the baseline. Because Pareto can also be ordered by likelihood ratio, optimal server allocation remains unchanged. Optimal buffer allocation requires simulation-based optimization techniques, and is found to be $b_2^* = 2, b_3^* = 3$.
- *Data Contamination*. he true service times for servers 1, 2, and 3 are the same as in the baseline case, with the optimal flowline design same as in the baseline. However, the data set for server 1's service

Table 4: Proportions of Incorrect Server Allocation

	Baseline		Lognormal		Contamination	
n	MHDE	MLE	MHDE	MLE	MHDE	MLE
25	1.18%	0.94%	2.24%	13.0%	3.84%	62.8%
50	0.40%	0.40%	0.16%	10.2%	1.06%	84.0%
100	0	0	0	7.92%	0.08%	92.4%

Table 5: Mean Squared Errors of b_2^*

	Baseline		Lognormal		Contamination	
n	MHDE	MLE	MHDE	MLE	MHDE	MLE
25	0.123	0.118	0.150	0.270	0.128	7.32E-2
50	5.30E-2	4.96E-2	7.30E-2	0.247	6.48E-2	1.44E-2
100	1.02E-2	9.20E-3	2.34E-2	0.218	1.84E-2	6.00E-4

Table 6: The Mean ROGs (in %) of the Throughput

	Baseline		Lognormal		Contamination	
n	MHDE	MLE	MHDE	MLE	MHDE	MLE
25	2.95	2.74	4.20	19.8	4.16	62.3
50	0.98	0.87	1.51	15.9	1.57	80.5
100	0.16	0.13	0.37	13.3	0.29	84.4

times is contaminated, modeled as a mixture of exponential distributions $0.9Exp(0.5)+0.1Exp(10)$. MLE would estimate $\widehat{\mu}_{1,n} \approx 1.45$, and thus would mistakenly place server 2 in the middle.

We first report the results on the proportion of times when server allocation is not correct, i.e., the fastest server 1 is not assigned to the middle station. Results are based on 5,000 i.i.d. replications. Epanechnikov kernel was used with the bandwidth c_n calculated using Silverman's plug-in bandwidth approach $c_n = 2.34 \times \widehat{\sigma} \times n^{-1/5}$, where $\widehat{\sigma}$ is the sample standard deviation of the data set.

We then report results on the MSE of the buffer space in front of station 2 in the following table

Finally, we report results on ROGs in Table 6. From the table, we see that MHDE consistently maintains the benefit of optimization using the estimated parameters for the exponential distribution model, despite the impact of model misspecification and data contamination. In contrast, data contamination has a drastic impact on the quality of the decision when MLE is used, losing up to 85% of the benefit of optimization.

6 CONCLUSION

In this paper, we propose to use MHDE instead of the commonly used MLE to estimate probability distribution models from data sets in stochastic optimization. We present results on the asymptotic properties of the estimated optimizers and optimal objective values, and demonstrated using two sets of experiments the efficiency and robustness of stochastic optimization with MHDE.

ACKNOWLEDGMENTS

Jie Xu's research is supported in part by the National Science Foundation under Grant No. CMMI- 1233376 and CMMI-1462787. Anand Vidyashankar's research was supported in part by a grant from NSF DMS 1107108.

REFERENCES

Barton, R. R., B. L. Nelson, and W. Xie. 2014. "Quantifying Input Uncertainty via Simulation Confidence Intervals". *INFORMS Journal on Computing* 26:74–87.

Ben-Tal, A., and A. Nemirovski. 1998. "Robust convex optimization". *Mathematics of Operations Research* 23 (4): 769–805.

Ben-Tal, A., and A. Nemirovski. 2000. "Robust solutions of linear programming problems contaminated with uncertain data". *Mathematical programming* 88 (3): 411–424.

Beran, R. 1977. "Minimum Hellinger distance estimates for parametric models". *The Annals of Statistics*:445–463.

Beran, R. 1978. "An efficient and robust adaptive estimator of location". *Ann. Statist.* 6 (2): 292–313.

Bertsimas, D., and M. Sim. 2004. "The price of robustness". *Operations research* 52 (1): 35–53.

Buzacott, J. A., and J. G. Shantikumar. 1993. *Stochastic Models of Manufacturing Systems*. Englewood Cliffs, NJ: Prentice-Hall.

Chan, S. S. 2008. *Robust and Efficient Inference for Linear Mixed Models using Skew-Normal Distributions*. Ph. D. thesis, Cornell University.

Cheng, A.-L., and A. Vidyashankar. 2006. "Minimum Hellinger distance estimation for randomized play the winner design". *Journal of Statistical Planning and Inference* 136:1875–1910.

Cheng, R., and W. Holland. 1997. "Sensitivity of computer simulation experiments to errors in input data". *Journal of Statistical Computation and Simulation* 58:219–241.

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

Delage, E., and Y. Ye. 2010. "Distributionally robust optimization under moment uncertainty with application to data-driven problems". *Operations Research* 58:595–612.

Dembo, A., and O. Zeitouni. 1998. *Large deviations techniques and applications*, Volume 2. Springer.

Fu, M. 1994. "Sample path derivatives for (s,S) inventory systems". *Operations Research* 42:351–364.

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

Glasserman, P. 1991. *Gradient Estimation via Perturbation Analysis*. Norwell, MA: Kluwer Academic Publisher.

Glasserman, P., and S. Tayur. 1995. "Sensitivity analysis for base-stock levels in multiechelon production-inventory systems". *Management Science* 41:263–281.

Ho, Y.-C., and X.-R. Cao. 1991. *Perturbation Analysis and Discrete Event Dynamic Systems*. Norwell, MA: Kluwer Academic Publisher.

Hong, L. J. 2009. "Estimating quantile sensitivities". *Operations Research* 57:118–130.

Hong, L. J., and G. Liu. 2009. "Simulating sensitivities of conditional value at risk". *Management Science* 55:281–293.

Hooker, G., and A. Vidyashankar. 2014. "Bayesian model robustness via disparities". *TEST* 23 (3): 556–584.

Hopp, W. J., and M. L. Spearman. 2011. *Factory physics*. Waveland Press.

Hu, Z., and L. J. Hong. 2013. "Kullback-Leibler divergence constrained distributionally robust optimization". Technical report, Hong Kong University of Science and Technology.

Kleywegt, A. J., A. Shapiro, and T. Homem-de Mello. 2002. "The sample average approximation method for stochastic discrete optimization". *SIAM Journal on Optimization* 12 (2): 479–502.

Lam, H. 2013a. "Robust sensitivity analysis for stochastic systems". Technical report, Department of Mathematics and Statistics, Boston University, Boston, MA.

Lam, H. 2013b. "Sensitivity to serial dependency of input processes: a robust approach". Technical report, Department of Mathematics and Statistics, Boston University, Boston, MA.

L'Ecuyer, P. 1990. "A unified view of the IPA, SF, and LR gradient estimation techniques". *Management Science* 36:1364–1383.

Lehmann, E. L., and G. Casella. 1998. *Theory of point estimation*, Volume 31. Springer.

McDiarmid, C. 1998. "Concentration". In *Probabilistic methods for algorithmic discrete mathematics*, 195–248. Springer.

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.

Ng, S.-H., and S. Chick. 2006. "Reducing parameter uncertainty for stochastic systems". *ACM Transactions on Modeling and Computer Simulation* 16:26–51.

Pichitlamken, J., and B. L. Nelson. 2003. "A combined procedure for optimization via simulation". *ACM Transactions on Modeling and Computer Simulation* 13:155–179.

Popescu, I. 2007. "Robust mean-covariance solutions for stochastic optimization". *Operations Research* 55 (1): 98–112.

Scarf, H., K. Arrow, and S. Karlin. 1958. "A min-max solution of an inventory problem". *Studies in the mathematical theory of inventory and production* 10:201–209.

Sriram, T., and A. Vidyashankar. 2000. "Minimum Hellinger distance estimation for supercritical Galton–Watson processes". *Statistics & probability letters* 50 (4): 331–342.

Vidyashankar, A., and J. Xu. 2015. "Distributionally robust stochastic optimization: a data-driven strategy". Technical report, George Mason University, Fairfax, VA.

Xie, W., B. Nelson, and R. Barton. 2013. "Statistical Uncertainty Analysis for Stochastic Simulation". Technical report, Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL.

Xie, W., B. Nelson, and R. Barton. 2014. "A Bayesian Framework for Quantifying Uncertainty in Stochastic Simulation". *Operations Research* forthcoming.

Xu, J., L. J. Hong, and B. L. Nelson. 2010. "Industrial Strength COMPASS: A Comprehensive Algorithm and Software for Optimization via Simulation". *ACM Transactions on Modeling and Computer Simulation* 20:3:1–3:29.

Yue, J., B. Chen, and M.-C. Wang. 2006. "Expected value of distribution information for the newsvendor problem". *Operations research* 54 (6): 1128–1136.

Zhu, Z., J. Zhang, and Y. Ye. 2006. "Newsvendor optimization with limited distribution information". Technical report, Working Paper, Stanford University, Stanford, CA.

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

Zouaoui, F., and J. R. Wilson. 2004. "Accounting for input-model and input-parameter uncertainties in simulation". *IIE Transactions* 36:1135–1151.

AUTHOR BIOGRAPHIES

ANAND N. VIDYASHANKAR is an Associate Professor in the Department of Statistics at George Mason University. He received his Ph.D. in Statistics and Mathematics from Iowa State University and has held subsequent positions in the Departments of Statistics at the University of Georgia and Cornell University. His main interests are in the areas of Branching Processes and Branching Random Walks, Stochastic Fixed Point Equations, Rare Event Simulations, Nested Simulations, Network Analysis, High-dimensional Statistical Inference, Robust Statistical Methods, and Risk Theory. He is a member of American Statistical Association, Institute of Mathematical statistics, and INFORMS. His email address is avidyash@gmu.edu.

JIE XU is an Assistant Professor in the Department of Systems Engineering and Operations Research at George Mason University. He received his Ph.D. from the Department of Industrial Engineering and Management Sciences of Northwestern University. His research interests include Monte Carlo simulation, stochastic optimization, computational intelligence, and applications in risk management and aviation. He is a member of INFORMS, IEEE, ACM, and SIAM. His email address is jxu13@gmu.edu.