Proceedings of the 2013 Winter Simulation Conference
R. Pasupathy, S.-H. Kim, A. Tolk, R. Hill, and M. E. Kuhl, eds.

ITERATIVE METHODS FOR ROBUST ESTIMATION UNDER BIVARIATE DISTRIBUTIONAL UNCERTAINTY

Henry Lam

Department of Mathematics and Statistics
Boston University
Boston, MA 02215, USA

Soumyadip Ghosh

Business Analytics and Math Sciences
T.J. Watson IBM Research Center
Yorktown Heights, NY 10598, USA

ABSTRACT

We propose an iterative algorithm to approximate the solution to an optimization problem that arises in estimating the value of a performance metric in a distributionally robust manner. The optimization formulation seeks to find a bivariate distribution that provides the worst-case estimate within a specified statistical distance from a nominal distribution and satisfies certain independence condition. This formulation is in general non-convex and no closed-form solution is known. We use recent results that characterize the local "sensitivity" of the estimation to the distribution used, and propose an iterative procedure on the space of probability distributions. We establish that the iterations of solutions are always feasible and that the sequence is provably improving the estimate. We describe conditions under which this sequence can be shown to converge to a locally optimal solution. Numerical experiments illustrate the effectiveness of this approach for a variety of nominal distributions.

1 INTRODUCTION

In many computational settings in engineering and management operations, the model assumed on the underlying system components is based on some convenient approximation of the truth. When the actual probability distribution that governs the model is different from the assumption, the decision output can potentially deviate to a substantial degree. This is commonly referred to as model uncertainty or ambiguity, and there is a sizeable literature devoted to studying this issue. For example, model selection methods based on classical and Bayesian statistics (Draper 1995; Chick 2001) and construction of input-adjusted confidence intervals have been proposed in large simulation studies. The latter can be done via delta-method approach (Zouaoui and Wilson 2004) or more sophisticated techniques, such as metamodel-assisted bootstrapping (Barton, Nelson, and Xie 2010; Barton 2012). In decision analysis, various forms of robust optimization have been introduced. The main streams of work include deterministic robust optimization that represents model ambiguity as deterministic uncertainty sets (Bertsimas, Brown, and Caramanis 2011; Ben-Tal and Nemirovski 2002), and distributionally robust optimization, where the underlying probability distributions are assumed to be ambiguous but come from a convenient family (Lim, Shanthikumar, and Shen 2006; Hu and Hong 2013). The latter approach is also particularly popular among the control community (Hansen and Sargent 2011), and recently, it has been applied to derive so-called robust Monte Carlo and has been used to compute robust bounds for performance measures that arise in the context of finance (Glasserman and Xu 2012).

In this paper, we consider a general form of robust estimation problems in the spirit of Glasserman and Xu (2012), which aims to generate bounds for performance measures that are subject to model uncertainty. Typically, these problems can be formulated as constrained maximization and minimization of the performance measure, say $E[h(\mathbf{X})]$, where $h(\cdot)$ is the *cost function* that is known (but not necessarily in closed-form) and $\mathbf{X}$ is the random component. The optimization is over the probability distribution P of $\mathbf{X}$

that generates the expectation $E[\cdot]$. Constraints are placed on the probability P to lie within an uncertainty set; one common example of such uncertainty set is a ball around a *benchmark* distribution, say, P_0, measured via a certain statistical distance, such as Chi-Square or Kullback-Leibler divergence. For instance, the cost function h can be the indicator of whether a portfolio with underlying stock price $\mathbf{X}$ hits a certain level of loss. While the process driving the stock is not known completely, portfolio managers typically have certain degree of confidence about the right distribution. The idea of the maximization program under the distribution constraint is to find the worst possible tail probability under such confidence. Similarly, the minimization program can be viewed as the best possible scenario. Depending on the application, these programs can have different interpretations and variations.

In many situations, a careful set-up of the constraints and the program formulation leads to tractable solution. For example, consider maximizing $E_f[h(X)]$ subject to a Kullback-Leibler constraint $D(P_f\|P_0) \leq \eta$, where D denotes the distance between the distribution P_f on X and the benchmark P_0, and η captures the level of confidence around the benchmark. This formulation has a neat explicit characterization of the optimal solution, which can be represented as an exponential change of measure on the cost $h(X)$ and an exponential parameter that is related to the confidence level η (Hansen and Sargent 2011). Analogous results have been shown to hold under other statistical distances, such as χ^2-distance (Chen and Lam 2013), with different forms of representation for the optimal changes of measure.

Our main focus of this paper is a natural scenario, especially in the context of simulation, where such tractable formulation breaks down. In particular, we consider restrictions that are placed on the *marginal* distributions of the random variate $\mathbf{X}$, defined as lower-dimensional (typically univariate) distributions that describe the probability law for sub-sets of components of $\mathbf{X}$. Previous work, such as Delage and Ye (2010) and Bertsimas, Popescu, and Sethuraman (2000), has considered marginal moment constraints in robust estimation formulation. In this article, we shall study marginal constraints based on χ^2-distance on a bivariate random vector that generates the objective function $E_f[h(X_1,X_2)]$, together with a natural constraint that the individual components X_i are i.i.d.. In other words, we restrict the space of feasible distributions P to those that are product form bivariate:

$$\begin{array}{ll} \max & E_f[h(X,Y)] \\ \text{subject to} & X,Y \sim P_f, i.i.d. \\ & \chi^2(P_f,P_0) \leq \eta \\ & P_f \in \mathscr{P}_0 \end{array} \tag{1}$$

(the min formulation is analogous) where the constraint is on the χ^2-distance between P_f and P_0, given by $\chi^2(P_f,P_0) = E_0(dP_f/dP_0 - 1)^2$, with dP_f/dP_0 being the Radon-Nikodym derivative of P_f with respect to P_0. The set $\mathscr{P}_0$ represents all P_f that are absolutely continuous with respect to P_0, a condition needed for χ^2-distance to be defined. The important feature here is that X and Y are assumed to be i.i.d. and they are both generated according to P_f, which is uncertain. The difficulty with the optimization problem (1) is that it is usually non-convex, and that no explicit form of the optimal solution is available.

Such formulation, in its more general form, appears commonly in stochastic systems, such as queueing networks, when many customers in the system each contribute as individual random sources and the associated vector of inputs (inter-arrival times and service-times) across the customers are i.i.d. The max and the min optimal values for the formulation type in (1) will together give a bounding interval for the performance measure subject to the specified constraints. As another motivating example, consider the following inventory model that will be studied in the numerical experiments section:

Example 1 A retailer has to decide on the size of order to place with a manufacturer that operates a batch-processing facility. The retailer requires that the order quantity q be sufficient to meet demand in two successive time periods, where each period's demand $d_i,\ i = 1,2$ is i.i.d. The distribution of d_i is determined to be P_0 by running the Chi-Square test statistic over available data, which returns an uncertainty estimate of $\chi^2(P,P_0) \leq \eta$. A profit of p units is realized on every sale, but the manufacturer charges a cost of c units for unsold inventory carried over at the end of each period. The net expected profit the retailer stands

to make under the nominal distribution P_0 is

$$E_0 h(d_1, d_2, q) = pE_0[\min\{d_1 + d_2, q\}] - cE_0[(q - d_1)^+] - cE_0[(q - d_1 - d_2)^+]. \tag{2}$$

Note that this decision problem is distinct from the classical newsvendor problem because of the presence of the extra carry-over cost charged at the end of the first period, represented in (2) by the middle term. Such formulations are typical of batch-produced goods such as steel slabs and hot-rolled coils from steel manufacturing plants.

We view (1) as an initial attempt to tackle more general robust estimation problems. With more involved analysis, our framework can be extended to the case where the random vector $\mathbf{X}$ has more than two i.i.d. components. It is also plausible that other statistical distances, such as Kullback Leibler divergence, can be used, as long as they possess a similar type of "local expansion" behavior that we shall discuss momentarily. Another generalization one might consider is the inclusion of moment constraints considered in, e.g., Delage and Ye (2010); we believe our method can be extended for this problem instance too.

The main goal of this paper is to derive a descent-type iterative procedure to find local optimum of (1). A tool that we shall utilize is the optimality characterization of (1) when η is chosen small enough, as a local expansion in terms of η, which was recently studied in Lam (2013) and Chen and Lam (2013). The result states that, for a bivariate cost function h, as $\eta \to 0$,

$$\max_{\chi^2(P_f, P_0) \le \eta} E_f[h(X,Y)] = E_0[h(X,Y)] + sd_0(H_0(X))\sqrt{\eta} + o(\eta) \tag{3}$$

where $H_0(x)$ is a "symmetrization" form of the cost function h, defined as

$$H_0(x) = E_0[h(X,Y)|X = x] + E_0[h(X,Y)|Y = x]$$

and $sd_0(\cdot)$ denotes the standard deviation under P_0. Note that $sd_0(\cdot)$ only acts on a single X in (3). This "symmetrization" arises in a "product rule" when differentiating the Lagrangian formulation in (1) (Lam 2013); alternatively, it can also be derived using a variation of Hoeffding decomposition (Serfling 2009). The key observation, in either method of derivation, is that on a *local* level, the optimal probability distribution within $\chi^2(P_f, P_0) \le \eta$ can be characterized by the change of measure

$$L^*(x) := \frac{dP_f}{dP_0}(x) = 1 + \frac{H_0(x) - E_0 H_0(X)}{sd_0(H_0(X))}\sqrt{\eta} \tag{4}$$

assuming that $sd_0(H_0(X)) > 0$ and η is small enough, under certain regularity conditions on h (such as boundedness, which we shall state in the sequel). In other words, the approximation (3) can be achieved by

$$E_0[h(X,Y)L^*(X)L^*(Y)] = E_0[h(X,Y)] + sd_0(H_0(X))\sqrt{\eta} + o(\eta) \tag{5}$$

We propose an iterative scheme whose local move is based on (5); Section 2 provides an outline of the proposed iterative procedure. In Section 3, we will prove some theoretical guarantees on the procedure. We show that our scheme is always feasible and ascending for the maximization formulation (1). Note that our proof does not need to know (3); it merely provides a guidance to find an ascending local move. Under mild boundedness conditions on the function h, we establish in Sub-section 3.2 that this sequence must then converge. Sub-section 3.3 provides a local optimality guarantee on the limit under stronger assumptions on the function h.

The iterative procedure described in Section 2 can be computed exactly for any benchmark distribution P_0 that has a discrete support. Section 4 describes a heuristic scheme to apply this procedure using a particle approach to the case where the support of the nominal distribution is continuous. Section 5 applies this heuristic to the inventory model in Example 1 under two assumptions of the benchmark P_0 and illustrates the efficiency of this approach.

2 THE ALGORITHM

We state an algorithm to approximate the solution of (1). Note that we can rewrite (1) in terms of the likelihood ratio $L := dP_f/dP_0$ as

$$\begin{array}{ll} \max & E_0[h(X,Y)L(X)L(Y)] \\ \text{subject to} & E_0(L-1)^2 \le \eta \\ & L \in \mathscr{L} \end{array} \tag{6}$$

where the χ^2-distance is expressed as $E_0(L-1)^2$, and $\mathscr{L} = \{L : E_0[L] = 1,\ L \ge 0 \text{ a.s.}\}$. It is easy to see that in general the program (6) is non-convex. As such, it is impossible to find a scheme that is guaranteed to approach the global optimum. We shall instead focus on finding the local optimum of (6). One heuristic to find the global optimum is to, for instance, run our algorithm at different initial points in the feasible set, in the space $\mathscr{L}$. Also, throughout our analysis we assume that h is bounded.

Suppose we start from the initial distribution P_0. At each step k, suppose that the current distribution is updated to P_k, and the distance from the benchmark is $D_k := \chi^2(P_k, P_0)$. Moreover, let $L_k := dP_k/dP_0$ and

$$H_k(x) = E_k[h(X,Y)|X = x] + E_k[h(X,Y)|Y = x]$$

be the "symmetrization" of the system under P_k, with $E_k[\cdot]$ being the expectation taken with respect to the distribution P_k. We also let $\bar{H}_k(x) = H_k(x) - E_k H_k(X)$ as the centered version of $H_k(x)$. Then, the following quantities are needed to define the next movement for P_{k+1}:

1. Define
$$A_k := -\frac{sd_k(H_k(X))^3}{2(E_k[h(X,Y)\bar{H}_k(X)\bar{H}_k(Y)])}$$
2. Define
$$B_k := \min\left\{-\frac{sd_k(H_k(X))}{\bar{H}_k(x)} : x \in \text{supp}(P_k) \text{ with } \bar{H}_k(x) < 0\right\}$$
3. Define
$$C_k := sd_k(H_k(X))\frac{\sqrt{(E_0[L_k(X)^2\bar{H}_k(X)])^2 + E_0[L_k(X)^2\bar{H}(X)^2](\eta - D_k)} - E_0[L_k(X)^2\bar{H}_k(X)]}{E_0[L_k(X)^2\bar{H}_k(X)^2]}$$

where $\text{supp}(P_k)$ denotes the support of P_k. We introduce a step size parameter δ_{k+1}, which defines the χ^2-distance that one should move in the next $(k+1)$-th step. To compute this parameter, we divide into two cases. Suppose that $H_k(X)$ is non-degenerate under P_k,

1. If $E_k[h(X,Y)\bar{H}_k(X)\bar{H}_k(Y)] < 0$, then
$$\sqrt{\delta_{k+1}} = \min\{A_k, B_k, C_k\}$$
2. If $E_k[h(X,Y)\bar{H}_k(X)\bar{H}_k(Y)] \ge 0$, then
$$\sqrt{\delta_{k+1}} = \min\{B_k, C_k\}$$

Next, we let

$$\nu_k(x) = \frac{\bar{H}_k(x)}{sd_k(H_k(X))}.$$

With δ_{k+1} and $\nu_k(\cdot)$, we define our change of measure at the $(k+1)$-th step as

$$L_{k+1}(x) = L_k(x)\left(1 + \nu_k(x)\sqrt{\delta_{k+1}}\right) \tag{7}$$

i.e. the probability distribution at the $(k+1)$-th step is

$$P_{k+1}(x) = P_k(x)\left(1 + v_k(x)\sqrt{\delta_{k+1}}\right) \tag{8}$$

This updates the probability distribution. After each iteration, the values of D_{k+1} and $H_{k+1}(x)$ are also updated. Repeat the procedure until either

1. $D_k = \eta$ or
2. $H_k(x)$ is degenerate under P_k.

We shall provide a few remarks on the algorithm. First, we explain some intuition on the purpose of the quantities A_k, B_k and C_k:

1. A_k ensures that the algorithm is strictly ascending from step k to $k+1$. As we shall see, the form of A_k comes from the analysis of a quadratic form that represents the increment from step k to $k+1$.
2. B_k ensures that the next probability distribution P_{k+1} is a valid distribution, i.e. the likelihood ratio that represents the change of measure at each step has to be non-negative.
3. C_k ensures that the next probability distribution P_{k+1} is feasible, i.e. $\chi^2(P_{k+1}, P_0) \leq \eta$.

We note that the form of the change of measure (7) resembles the best local movement as in (5). The parameter δ_k captures the confidence level such that the linear approximation in (3) holds. If δ_k is chosen small enough, we can then argue that the movement is ascending in the objective value. The quantities A_k, B_k and C_k combine to provide such guarantee.

In the next section, we will show rigorously that the algorithm is always strictly ascending, until one of the two stopping criteria holds. The first criterion means that a boundary point of the set $\chi^2(P_f, P_0) \leq \eta$ is reached. The second criterion, roughly speaking, is a condition for reaching a "stationary point". We will discuss in Sub-section 3.3 a semi-definite condition that guarantees that this "stationary point" is a local maximum in certain sense (the definition of local maximum is not entirely obvious, since χ^2-distance is not a metric).

3 THEORETICAL PROPERTIES

In this section we discuss a few properties of our algorithm. These include the strict ascendency of the algorithm, an optimality condition on the end value of the algorithm, and a convergence property of the "derivative" term $sd_k(H_k(X))$ in (5) in a sense that we shall discuss.

3.1 Ascendency Guarantee

The first property can be summarized as follows:

Theorem 1 Suppose the cost function h is bounded. The algorithm in Section 2 is always feasible, i.e. $\chi^2(P_k, P_0) \leq \eta$, and strictly ascending, i.e. $E_{k+1}[h(X,Y)] > E_k[h(X,Y)]$, until one of the two stopping criteria is met.

Proof. Note that $\sqrt{\delta_{k+1}}$ is the minimum of either the three terms A_k, B_k and C_k or the two terms B_k and C_k, under the corresponding conditions in Section 2. Related to the discussion before, we shall show that

1. A_k ensures that the algorithm is strictly ascending from step k to $k+1$.
2. B_k ensures that the next probability distribution P_{k+1} is a valid distribution.
3. C_k ensures that the next probability distribution P_{k+1} is feasible.

We shall show these one by one:

Ascendency. First, by choosing $\sqrt{\delta_{k+1}} \leq A_k$, we have

$$\begin{aligned}
E_{k+1}[h(X,Y)] &= E_k\left[h(X,Y)\left(1+\frac{\bar{H}_k(X)}{sd_k(H_k(X))}\sqrt{\delta_{k+1}}\right)\left(1+\frac{\bar{H}_k(Y)}{sd_k(H_k(Y))}\sqrt{\delta_{k+1}}\right)\right] \\
&= E_k[h(X,Y)] + \frac{E_k[h(X,Y)(\bar{H}_k(X)+\bar{H}_k(Y))]}{sd_k(H_k(X))}\sqrt{\delta_{k+1}} \\
&\quad + \frac{E_k[h(X,Y)\bar{H}_k(X)\bar{H}_k(Y)]}{Var_k(H_k(X))}\delta_{k+1} \\
&= E_k[h(X,Y)] + \frac{E_k[E_k[h(X,Y)|X]\bar{H}_k(X)] + E_k[E_k[h(X,Y)|Y]\bar{H}_k(Y)]}{sd_k(H_k(X))}\sqrt{\delta_{k+1}} \\
&\quad + \frac{E_k[h(X,Y)\bar{H}_k(X)\bar{H}_k(Y)]}{Var_k(H_k(X))}\delta_{k+1} \\
&= E_k[h(X,Y)] + sd_k(H_k(X))\sqrt{\delta_{k+1}} + \frac{E_k[h(X,Y)\bar{H}_k(X)\bar{H}_k(Y)]}{Var_k(H_k(X))}\delta_{k+1} \qquad (9)
\end{aligned}$$

by using the i.i.d. assumption on X and Y in the last equality. Suppose $E_k[h(X,Y)\bar{H}_k(X)\bar{H}_k(Y)] \geq 0$, then picking a positive δ_{t+1} will make the algorithm ascending; if $E_k[h(X,Y)\bar{H}_k(X)\bar{H}_k(Y)] < 0$, then $\sqrt{\delta_{k+1}} \leq A_k$ guarantees that

$$sd_k(H_k(X))\sqrt{\delta_{k+1}} + \frac{E_k[h(X,Y)\bar{H}_k(X)\bar{H}_k(Y)]}{Var_k(H_k(X))}\delta_{k+1} \geq 0.$$

In fact, picking

$$\sqrt{\delta_{k+1}} = -\frac{sd_k(H_k(X))^3}{2E_k[h(X,Y)\bar{H}_k(X)\bar{H}_k(Y)]}$$

maximizes the increase in $E_k[h(X,Y)]$. Hence the algorithm is strictly ascending.

Validity of probability distribution. Note that we have defined in (8) that

$$\frac{dP_{k+1}}{dP_k} = 1 + \nu_k(x)\sqrt{\delta_{k+1}}.$$

To guarantee that P_{k+1} is valid, we must have $E_k[dP_{k+1}/dP_k] = 1$, which is easily seen to hold since $E_k[1+\nu_k(X)\sqrt{\delta_{k+1}}] = 1 + E_k[\nu_k(X)]\sqrt{\delta_{k+1}} = 1$. Secondly, we also need δ_{k+1} to satisfy the constraint

$$1 + \nu_k(x)\sqrt{\delta_{k+1}} \geq 0$$

for all $x \in \text{supp}\ (P_k)$. We need only scrutinize the x that have $\nu_k(x) < 0$, or $\bar{H}_k(x) < 0$, and make sure that

$$\sqrt{\delta_{k+1}} \leq -\frac{1}{\nu_k(x)}$$

for all such x. This gives B_k.

Feasibility. Finally, we need to ensure that $D_{k+1} = \chi^2(P_{k+1}, P_0) \leq \eta$. Note that

$$\begin{aligned}
D_{k+1} = \chi^2(P_{k+1}, P_0) &= E_0[L_k(X)^2(1+\nu_k(x)\sqrt{\delta_{k+1}})^2] - 1 \\
&= E_0[L_k(X)^2 + 2L_k(X)^2\nu_k(X)\sqrt{\delta_{k+1}} + L_k(X)^2\nu_k(X)^2\delta_{k+1}] - 1 \\
&= E_0[L_k(X)^2] - 1 + 2E_0[L_k(X)^2\nu_k(X)]\sqrt{\delta_{k+1}} + E_0[L_k(X)^2\nu_k(X)^2]\delta_{k+1} \\
&= D_k + 2E_0[L_k(X)^2\nu_k(X)]\sqrt{\delta_{k+1}} + E_0[L_k(X)^2\nu_k(X)^2]\delta_{k+1}.
\end{aligned}$$

Hence we need

$$D_k + 2E_0[L_k(X)^2 \nu_k(X)]\sqrt{\delta_{k+1}} + E_0[L_k(X)^2 \nu_k(X)^2]\delta_{k+1} \leq \eta.$$

Since this is quadratic and convex in $\sqrt{\delta_{k+1}}$, we need that

$$\sqrt{\delta_{k+1}} \leq \xi$$

where ξ is the positive root of the equation

$$D_k + 2E_0[L_k(X)^2 \nu_k(X)]\sqrt{\delta_{k+1}} + E_0[L_k(X)^2 \nu_k(X)^2]\delta_{k+1} = \eta.$$

This gives

$$\sqrt{\delta_{k+1}} \leq \frac{\sqrt{(E_0[L_k(X)^2 \nu_k(X)])^2 + E_0[L_k(X)^2 \nu_k(X)^2](\eta - D_k)} - E_0[L_k(X)^2 \nu_k(X)]}{E_0[L_k(X)^2 \nu_k(X)^2]}.$$

Note that the right hand side above is C_k. This will ensure that $D_{k+1} \leq \eta$.

□

3.2 Convergence

We shall also discuss a convergence property of the algorithm. Let us focus on how the algorithm behaves in the interior points of the set $\chi^2(P_f, P_0) \leq \eta$. In particular, we have the following:

Proposition 1 Consider a bounded cost function h. Suppose $\eta = \infty$, i.e. the constraint $\chi^2(P_f, P_0) \leq \eta$ is removed from (1). Then our algorithm will converge to a probability distribution P^* that satisfies $Var^*(H^*(X)) = 0$ (where $Var^*(\cdot)$ and $H^*(\cdot)$ are defined as the variance and symmetrization of h under P^*).

Proof. Note that in this case, the quantity C_k in our algorithm is not needed. If $E_k[h(X,Y)\bar{H}_k(X)\bar{H}_k(Y)] < 0$, then $\sqrt{\delta_{k+1}} = \min\{A_k, B_k\}$, otherwise $\sqrt{\delta_{k+1}} = B_k$.

We shall prove by contradiction. Suppose that the algorithm does not lead the quantity $Var_k(H_k(X))$ to 0, then there must exist a subsequence $\{k_i\}_{i=1,2,\ldots}$ such that $Var_{k_i}(H_{k_i}(X)) > \varepsilon$ for all i for some $\varepsilon > 0$.

Now, from the calculation in (9), we know the increment $E_{k+1}[h(X,Y)] - E_k[h(X,Y)]$ is

$$sd_k(H_k(X))\sqrt{\delta_{k+1}} + \frac{E_k[h(X,Y)\bar{H}_k(X)\bar{H}_k(Y)]}{Var_k(H_k(X))}\delta_{k+1}. \tag{10}$$

Consider the step k_i, and we shall analyze (10) explicitly. We shall divide into three cases:

1. $E_{k_i}[h(X,Y)\bar{H}_{k_i}(X)\bar{H}_{k_i}(Y)] < 0$ *and* $\sqrt{\delta_{k_i+1}} = A_{k_i}$*:* We have (10) equal to

$$\begin{aligned} & sd_{k_i}(H_{k_i}(X))A_{k_i} + \frac{E_{k_i}[h(X,Y)\bar{H}_{k_i}(X)\bar{H}_{k_i}(Y)]}{Var_{k_i}(H_{k_i}(X))}A_{k_i}^2 \\ = \; & sd_{k_i}(H_{k_i}(X))\left(-\frac{sd_{k_i}(H_{k_i}(X))^3}{2E_{k_i}[h(X,Y)\bar{H}_{k_i}(X)\bar{H}_{k_i}(Y)]}\right) \\ & + \frac{E_{k_i}[h(X,Y)\bar{H}_{k_i}(X)\bar{H}_{k_i}(Y)]}{Var_{k_i}(H_k(X))}\left(-\frac{sd_{k_i}(H_{k_i}(X))^3}{2E_{k_i}[h(X,Y)\bar{H}_{k_i}(X)\bar{H}_{k_i}(Y)]}\right)^2 \\ = \; & -\frac{Var_{k_i}(H_{k_i}(X))^2}{4E_{k_i}[h(X,Y)\bar{H}_{k_i}(X)\bar{H}_{k_i}(Y)]} \end{aligned} \tag{11}$$

But since h is bounded, so is $-E_{k_i}[h(X,Y)\bar{H}_{k_i}(X)\bar{H}_{k_i}(Y)]$, and since we assume $Var_{k_i}(H_{k_i}(X)) > \varepsilon$, we have (11) bounded away from 0 (and positive).

2. $E_{k_i}[h(X,Y)\bar{H}_{k_i}(X)\bar{H}_{k_i}(Y)] < 0$ *and* $\sqrt{\delta_{k_i+1}} = B_{k_i}$*:* Let $\bar{H}^*_k := \text{ess inf}_{x\in\text{supp}(P_k):\bar{H}_k(x)<0}\bar{H}_k(x) < 0$. In this case, since $A_{k_i} \geq B_{k_i}$, we have

$$-\frac{sd_{k_i}(H_{k_i}(X))^3}{2E_{k_i}[h(X,Y)\bar{H}_{k_i}(X)\bar{H}_{k_i}(Y)]} \geq -\frac{sd_{k_i}(H_{k_i}(X))}{\bar{H}^*_{k_i}}$$

which gives

$$-E_{k_i}[h(X,Y)\bar{H}_{k_i}(X)\bar{H}_{k_i}(Y)] \leq -\frac{\bar{H}^*_{k_i}Var_{k_i}(H_{k_i}(X))}{2} \tag{12}$$

Note that (10) is equal to

$$\begin{aligned} & sd_{k_i}(H_{k_i}(X))B_{k_i} + \frac{E_{k_i}[h(X,Y)\bar{H}_{k_i}(X)\bar{H}_{k_i}(Y)]}{Var_{k_i}(H_{k_i}(X))}B^2_{k_i} \\ = \;& sd_{k_i}(H_{k_i}(X))\left(-\frac{sd_{k_i}(H_{k_i}(X))}{\bar{H}^*_{k_i}}\right) + \frac{E_{k_i}[h(X,Y)\bar{H}_{k_i}(X)\bar{H}_{k_i}(Y)]}{Var_{k_i}(H_{k_i}(X))}\frac{Var_{k_i}(H_{k_i}(X))}{(\bar{H}^*_{k_i})^2} \\ = \;& -\frac{Var_{k_i}(H_{k_i}(X))}{\bar{H}^*_{k_i}} + \frac{E_{k_i}[h(X,Y)\bar{H}_{k_i}(X)\bar{H}_{k_i}(Y)]}{(\bar{H}^*_{k_i})^2} \\ \geq \;& -\frac{Var_{k_i}(H_{k_i}(X))}{2\bar{H}^*_{k_i}} \end{aligned} \tag{13}$$

where the last inequality follows from (12). Since $Var_{k_i}(H_{k_i}) > \varepsilon$ and $\bar{H}^*_{k_i}$ is bounded from above since h is bounded, we have (13) bounded away from 0.

3. $E_{k_i}[h(X,Y)\bar{H}_{k_i}(X)\bar{H}_{k_i}(Y)] \geq 0$*:* In this case we have $\sqrt{\delta_{k_i+1}} = B_{k_i}$. Similar to above, (10) is equal to

$$\begin{aligned} & sd_{k_i}(H_{k_i}(X))B_{k_i} + \frac{E_{k_i}[h(X,Y)\bar{H}_{k_i}(X)\bar{H}_{k_i}(Y)]}{Var_{k_i}(H_{k_i}(X))}B^2_{k_i} \\ = \;& -\frac{Var_{k_i}(H_{k_i}(X))}{\bar{H}^*_{k_i}} + \frac{E_{k_i}[h(X,Y)\bar{H}_{k_i}(X)\bar{H}_{k_i}(Y)]}{(\bar{H}^*_{k_i})^2} \end{aligned} \tag{14}$$

Since in this case $E_{k_i}[h(X,Y)\bar{H}_{k_i}(X)\bar{H}_{k_i}(Y)] \geq 0$, it follows similarly that (14) is bounded away from 0.

In conclusion, for every step in the sequence k_i, the increment $E_{k_i+1}[h(X,Y)] - E_{k_i}[h(X,Y)]$ is always bounded away from 0. Hence $E_{k_i}[h(X,Y)]$ grows in an unbounded manner. But we have assumed that h is bounded, hence a contradiction. □

3.3 Nesting and Optimality

In this sub-section we shall discuss a simple optimality condition of our algorithm. Before so, we emphasize that the use of χ^2-distance as a measurement of disparateness between distributions, our decision variables, poses some non-standard features to the definition of optimality. At the foremost, χ^2-distance is not a proper metric: it does not satisfy triangle inequality nor commutativity. This itself does not pose direct issue for using iterative schemes. However, it leads to a "one-way" update of P_k, or a nesting behavior in terms of absolute continuity. Namely, the sequence of update $\{P_k\}_{k=0,1,2,\ldots}$ satisfies $P_{k+1} \prec P_k$, where $\prec$ denotes "absolutely continuous with respect to". In other words, the support of P_k shrinks at each step.

We shall call P^* a *nested locally optimal* distribution if for any $P \prec P^*$, we have $E[h(X,Y)] \leq E^*[h(X,Y)]$. We call P^* a *nested strict locally optimal* distribution if the inequality is strict. The next proposition gives a condition for local optimality, if the algorithm stops at an interior point of the feasible region. It requires a semi-definite property on the cost function h.

Proposition 2 Consider a probability distribution P^* that satisfies $Var^*(H^*(X)) = 0$. If $h(x,y)$ is negative semi-definite in $\text{supp}(P^*)$, in the sense that

$$\int h(x,y)d\varepsilon(x)d\varepsilon(y) \leq 0 \tag{15}$$

for any signed measure ε on $\text{supp}(P^*)$ with bounded variation, then P^* is a nested local maximum of the program (1). Supposing that $h(x,y)$ is negative definite, i.e. the inequality (15) is strict for any non-zero measure, then P^* is a nested strict local maximum.

Proof. Suppose P^* satisfies $Var^*(H^*(X)) = 0$. Consider any $P \prec P^*$ in a neighborhood of P^*. The proof follows from the following decomposition (with some abuse of notation)

$$\begin{aligned}
E[h(X,Y)] &= \int h(x,y)dP(x)dP(y) \\
&= \int h(x,y)d(P^*(x)+(P-P^*)(x))d(P^*(y)+(P-P^*)(y)) \\
&= \int h(x,y)dP^*(x)dP^*(y) + \int h(x,y)dP^*(x)d(P-P^*)(y) + \int h(x,y)dP^*(y)d(P-P^*)(x) \\
&\quad + \int h(x,y)d(P-P^*)(x)d(P-P^*)(y) \\
&= E^*[h(X,Y)] + \int H^*(x)d(P-P^*)(x) + \int h(x,y)d(P-P^*)(x)d(P-P^*)(y)
\end{aligned}$$

where the last equality follows since x and y are dummies for the variable X. Hence if $Var^*(H^*(X)) = 0$, then $H^*(X)$ is degenerate and the second term above vanishes. If h is negative semi-definite in $\text{supp}(P^*)$, then we have $E[h(X,Y)] \leq E^*[h(X,Y)]$, and hence P^* is a nested local maximum. Similarly, if h is negative definite, then P^* is a nested strict local maximum. □

Note that the proof above resembles the optimality conditions in Euclidean space, the main difference now being that we are working on the space of distributions. The symmetrization $H^*(x)$ can be interpreted as the "derivative" in such space (see also Lam (2013)), and the semi-definite condition is also analogous.

4 SIMULATION HEURISTIC FOR CONTINUOUS P_0

This section describes an algorithm for efficiently calculating the distributions implied by the iterative procedure described above. The calculations are exact for benchmark distributions that have finite discrete support. For the case when the support of the benchmark distribution P_0 is continuous, we suggest a naïve approximation procedure of approximating the original benchmark P_0 with a discrete distribution $\hat{P}_0$ that equi-weighs a set of N_0 samples generated from P_0 in an i.i.d. fashion.

HEURISTIC 1: ROBUST ESTIMATION APPROXIMATION ALGORITHM

Given: benchmark distribution P_0, distribution χ^2-discrepancy target η, oracle to measure performance metric $h(x_1,x_2)$

1. If benchmark P_0 has support $\mathscr{X}$ that is discrete, set $\hat{\mathscr{X}} = \mathscr{X}$. Else, set $\hat{\mathscr{X}} = \{x_i, i = 1,\ldots,N_0\}$, where each x_i is sampled from the benchmark distribution P_0. Let $\hat{P}_0 = \{1/N_0,\ldots,1/N_0\}$ be an N_0-dimensional row-vector that represents the approximate benchmark distribution over support-set $\hat{\mathscr{X}}$. $L_0 = \{1,\ldots,1\}$, a N_0-dimensional row-vector that records the likelihood ratio $\hat{P}_k/\hat{P}_0$. Record $\mathbf{h}_0 = \{h(x_1,x_j),\ \forall x_i,x_i \in \hat{\mathscr{X}}\}$, the matrix of values taken by the performance oracle h over $\hat{\mathscr{X}} \times \hat{\mathscr{X}}$.
2. For $k = 0,1,2,\ldots$

(a) Calculate $E_k h = \hat{P}_k \mathbf{h}_0 \hat{P}_k^t$.
(b) Calculate $D_k = \chi^2(\hat{P}_k, \hat{P}_0) = \sum_i^{N_0} (L_k(i) - 1)^2 * \hat{P}_k(i)$. If $D_k = \eta$, proceed to Step 3.
(c) Calculate $H_k = (\mathbf{h}_0 + \mathbf{h}_0^t)\hat{P}_k$, the symmetric measure under the kth density iterate.
(d) Calculate $E_k H_k$ and $sd_k(H_k)$, the mean and standard deviation of the symmetric measure under $\hat{P}_k$. If $sd_k(H_k) = 0$, then proceed to Step 3.
(e) Calculate the three limits A_k, B_k and C_k on the distributional distance δ_k as described in Section 2. Set $\delta_k = \min\{A_k, B_k, C_k\}$. If $\delta_k = 0$, proceed to Step 3.
(f) Set $L_{k+1}(i) = L_k(i) * \left(1 + \frac{H_k(i) - E_k H_k}{sd_k(H_k)} \sqrt{\delta_k}\right)$, $\forall i = 1, \ldots, N_0$.
(g) Set $\hat{P}_{k+1}(i) = \hat{P}_0(i) * L_{k+1}(i)$, $\forall i = 1, \ldots, N_0$.

3. Return $\hat{P}_k$ and $E_k h$ as estimates of the desired worst-case density and the performance metric.

5 NUMERICAL RESULTS

This section analyzes Example 1 that models a retailer of batch-produced goods. Recall that the expected profit the retailer stands to make under the nominal distribution P_0 is given by (2) to be:

$$E_0 h(d_1, d_2, q) = pE_0[\min\{d_1 + d_2, q\}] - cE_0[(q - d_1)^+] - cE_0[(q - d_1 - d_2)^+]. \tag{16}$$

A retailer has to decide on the size of the order q to place with a manufacturer that operates a batch-processing facility. The classical approach to this problem is to first determine the benchmark distribution P_0 from available data, and then set the order quantity to be the optimizer of the stochastic optimization problem that maximizes the realized profit:

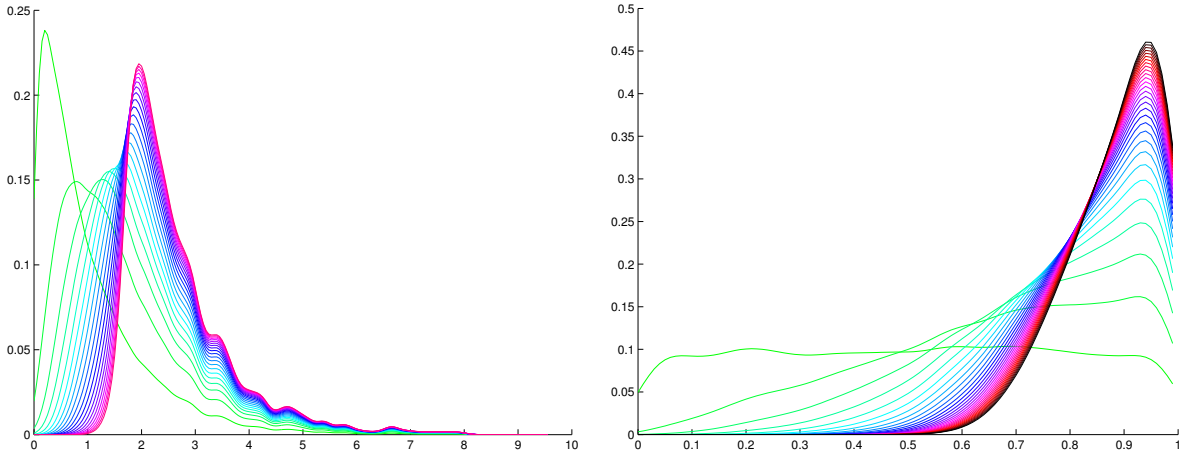
$$q^* = argmin_q \, E_0 h(d_1, d_2, q). \tag{17}$$

Differentiating (16) with respect to q gives us the following optimization criterion:

$$\frac{\partial E_0 h(d_1, d_2, q)}{\partial q} = p - cP_0(q) - (p + c) \int_{-\infty}^{\infty} P_0(q - x) dP_0(x) = 0.$$

where we abuse notation to denote P_0 also as the (identical) distribution function of d_1 and d_2. Suppose the nominal distribution P_0 is $exp(\lambda)$ and $p = rc, r > 1$. Then, the optimal q^* is the unique solution to the equation $e^{-\lambda q}(r + 2 + \lambda(r+1)q) = 2$. Taking the second derivative of (16) with respect to q and setting $q = q^*$ confirms that this is a maximizer. For instance, if $\lambda = 1$ and $r = 3$, then $q^* = 1.812$. As remarked earlier, the problem (17) is distinct from the newsvendor model because of the presence of the additional first-period cost term in (16). In its absence, the optimal order quantity from the newsvendor model would be the $p/(p+c)$-th quantile of the distribution of the sum $(d_1 + d_2)$. In the case where $p/c = r = 3$ and $P_0 \sim exp(1)$, $(d_1 + d_2)$ is an Erlang distribution, and the optimal newsvendor order quantity is 2.6926. Thus, the optimal order quantity falls to 1.812 because of the additional carryover cost in the first period.

We apply our algorithm in Section 2. Since the exponential distribution P_0 is continuous, we use Heuristic 1 in Section 4 to encode the distribution. For illustrative purpose, we keep track of the discrepancy distance $\chi^2(P_k, P_0)$ instead of putting a hard constraint that it is bounded by η (when the constraint is imposed, the algorithm will stop whenever the η-boundary is reached). By Proposition 1, we know that eventually the algorithm should stop when $sd_k(H_k(X)) = 0$. Figure 1(a) plots the densities of the P_k iterates generated by applying Heuristic 1 until $sd_k(H_k(X))$ is within a small tolerance around 0. The initial sample size was set to $N_0 = 10,000$. The performance measure (16) rises from $E_0[h(X,Y)] = 5.941$ to the optimal $E^*[h(X,Y)] = 9.076$, where $\chi^2(P^*, P_0) = 0.33$. (Figure 1(a) plots a smoothed version of the empirical density maintained by Heuristic 1, where the smoothing is obtained by binning the N_0 support points into 50 equi-sized bins and then constructing a Gaussian kernel density with an appropriate smoothing

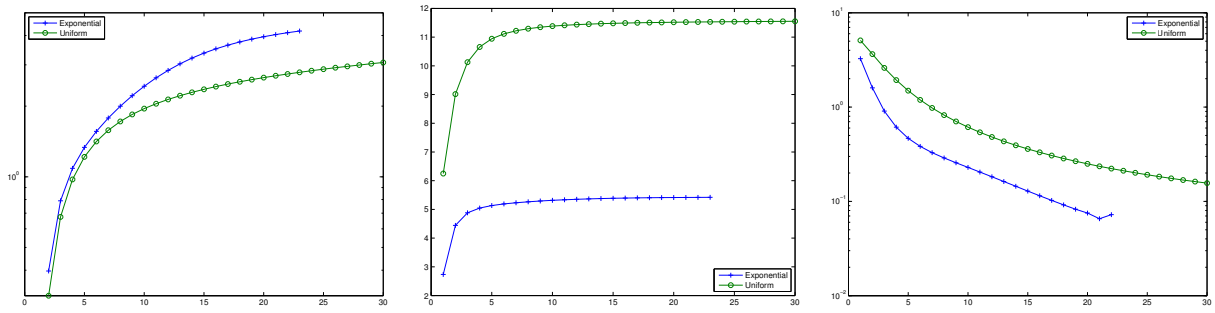


(a) $P_0 \sim exp(1)$ (b) $P_0 \sim U(0,1)$

Figure 1: Densities Generated by Iterative Heuristic 1, Smoothed Using Gaussian Kernels.

parameter. This smoothing is meant as a visual aid and has no bearing on the calculations in Heuristic 1.) Figure 1(b) plots the densities of the iterates for the same problem but with $P_0 \sim U[0,1]$, $q* = 1.524$ and $r = 8$.

Of perhaps more interest is Figure 2 that plots the objective function $E_k[h(X,Y)]$ in (2), the discrepancy measure $\chi^2(P_k,P_0)$ and the "gradient" $sd_k(H_k(X))$ associated with the k-th iteration of Heuristic 1 for both the exponential and the uniform benchmark cases. The results show that Heuristic 1 increasingly tends to concentrate the mass in a subset of the original sample-set of N_0 points that reduces the value of $sd_k(H_k(X))$. In the exponential-P_0 case, the method was able to identify a distribution with $sd_k(H_k(X)) = 0$ within $k = 22$ iterations. The identified solution is however only an approximation of the true optimal density (when no discrepancy constraint is imposed), since the starting P_0 used in Heuristic 1 is a sample set generated from the original benchmark distribution.



(a) $\chi^2(P_k,P_0)$ (b) Performance Metric $E_k h(X,Y)$ (c) $sd_k(H_k(X))$

Figure 2: Performance of Heuristic 1 as Iteration Count Grows, in the Exponential and Uniform Benchmark Cases.

REFERENCES

Barton, R. R. 2012. "Tutorial: Input uncertainty in output analysis". In *Proceedings of the 2012 Winter Simulation Conference*, edited by C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A. M. Uhrmacher, 1–12. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Barton, R. R., B. L. Nelson, and W. Xie. 2010. "A framework for input uncertainty analysis". In *Proceedings of the 2010 Winter Simulation Conference*, edited by B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yucesan, 1189–1198. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Ben-Tal, A., and A. Nemirovski. 2002. "Robust optimization–methodology and applications". *Mathematical Programming* 92 (3): 453–480.

Bertsimas, D., D. B. Brown, and C. Caramanis. 2011. "Theory and applications of robust optimization". *SIAM Review* 53 (3): 464–501.

Bertsimas, D., I. Popescu, and J. Sethuraman. 2000. "Moment problems and semidefinite optimization". In *Handbook of semidefinite programming*, 469–509. Springer.

Chen, X., and H. Lam. 2013. "Robust Assessment of Model Uncertainty for Steady-State Estimators". *Working Paper*.

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

Delage, E., and Y. Ye. 2010. "Distributionally robust optimization under moment uncertainty with application to data-driven problems". *Operations Research* 58 (3): 595–612.

Draper, D. 1995. "Assessment and propagation of model uncertainty". *Journal of the Royal Statistical Society. Series B (Methodological)*:45–97.

Glasserman, P., and X. Xu. 2012. "Robust Risk Measurement and Model Risk". *Available at SSRN 2167765*.

Hansen, L. P., and T. J. Sargent. 2011. *Robustness*. Princeton university press.

Hu, Z., and J. Hong. 2013. "Kullback-Leibler Divergence Constrained Distributionally Robust Optimization". *Working Paper*.

Lam, H. 2013. "Robust Sensitivity Analysis for Stochastic Systems". *arXiv preprint arXiv:1303.0326*.

Lim, A. E., J. G. Shanthikumar, and Z. M. Shen. 2006. "Model uncertainty, robust optimization, and learning". *Tutorials in Operations Research: Models, Methods, and Applications for Innovative Decision Making*:66–94.

Serfling, R. J. 2009. *Approximation theorems of mathematical statistics*, Volume 162. Wiley-Interscience.

Zouaoui, F., and J. R. Wilson. 2004. "Accounting for input-model and input-parameter uncertainties in simulation". *IIE Transactions* 36 (11): 1135–1151.

AUTHOR BIOGRAPHIES

HENRY LAM is an Assistant Professor in the Department of Mathematics and Statistics at Boston University. He graduated from Harvard University with a Ph.D. degree in statistics in 2011. His research interests lie in applied probability and Monte Carlo methods with applications in queueing, operations management and insurance modeling. His email is khlam@bu.edu and his web page is at http://math.bu.edu/people/khlam/.

SOUMYADIP GHOSH is a Research Staff Member in the Business Analytics and Mathematical Sciences Department at the IBM T.J. Watson Research Center. His current research interests lie in simulation based optimization techniques for stochastic optimization problems, with a focus on applications in Energy and Power systems and supply chain management. His email is ghoshs@us.ibm.com and his web page is at https://researcher.ibm.com/researcher/view.php?person=us-ghoshs.