Proceedings of the 2002 Winter Simulation Conference
E. Yücesan, C.-H. Chen, J. L. Snowdon, and J. M. Charnes, eds.

ESTIMATION OF RARE EVENT PROBABILITIES USING CROSS-ENTROPY

Tito Homem-de-Mello

Department of IWSE
Ohio State University
Columbus, OH 43210-1271, U.S.A.

Reuven Y. Rubinstein

Faculty of Industrial Engineering and Management
Technion—Israel Institute of Technology
Haifa 32000, ISRAEL

ABSTRACT

This paper deals with estimation of probabilities of rare events in static simulation models using a fast adaptive two-stage procedure based on importance sampling and Kullback-Liebler's cross-entropy (CE). More specifically, at the first stage we estimate the optimal parameter vector in the importance sampling distribution using CE, and at the second stage we estimate the desired rare event probability using importance sampling (likelihood ratios). Some theoretical aspects of the proposed method, including its convergence, are established. The numerical results presented suggest that the method effectively estimates rare event probabilities.

1 INTRODUCTION

The performance of computer and communications systems is often characterized by the probability of certain *rare events* and it is frequently studied through simulation. A typical example is the probability of failure of a certain network, which is a measure of the reliability of that system. The use of crude Monte Carlo techniques, however, requires a prohibitively large numbers of trials in most interesting cases, so new techniques are required. Among the methods developed are the *splitting/RESTART* (see, for instance, Garvels and Kroese 1998; Glasserman et al. 1999; Görg 1999; Villén-Altamirano and Villén-Altamirano 1999) and *importance sampling* techniques (see, e.g., Glynn and Iglehart 1989).

The main idea of importance sampling (IS), when applied to rare events, is to make their occurrence more frequent, or in other words, to "speed up" the simulation, and at the same time keep the variance under control. It is well-known that, in theory, there exists a change of measure that yields *zero variance* estimators. Such optimal measure, however, typically cannot be computed exactly since it depends on the underlying quantities being estimated. One approach to find the right change of measure, appropriate for smaller systems, is described by results based on large deviations theory; see Asmussen and Rubinstein (1995), Heidelberger (1995), Kovalenko (1995), and Shahabuddin (1995) for surveys.

Another approach to the above problem can be derived when the underlying distribution belongs to some *parametric* family. We can then constrain the choice of IS distributions to the same family. Although such approach does not give the optimal zero-variance measure, it typically yields significant variance reduction; see, for instance, Rubinstein and Melamed (1998), Rubinstein and Shapiro (1993). On the other hand, the advantage of such procedure is that the resulting variance-minimization problem is finite-dimensional and as such can be tackled with optimization techniques. Still, the problem can be difficult to solve, since it is a stochastic optimization problem which is, in general, nonconvex. In Rubinstein (1997), an *adaptive* IS algorithm for rare events simulation was proposed in which the change of measure is *estimated* by minimizing the sample variance of the IS estimator.

An alternative to the variance minimization approach is to find the parameter that minimizes the "distance" between the IS distribution and the (unknown) optimal zero-variance measure. One particular distance function that has been proven useful is the so-called Kullback-Liebler's cross-entropy. A major advantage of such approach is that the resulting optimization problems are well-structured; indeed, in some cases they can be solved analytically. Moreover, as the events become rarer, the obtained parameter tends to coincide with the parameter that minimizes variance. This approach has been used in connection with combinatorial optimization problems, see Rubinstein (1999), de Boer et al. (2001).

In this paper we concentrate on the application of the cross-entropy method (henceforth called CE method) to estimate rare event probabilities in *static* models. We present an algorithm, discuss its convergence, and present some numerical results. An expanded discussion can be found in Homem-de Mello and Rubinstein (2002). An application of the CE method to dynamic systems such as queueing networks is given in de Boer, Kroese, and Rubinstein (2001).

2 BACKGROUND ON IMPORTANCE SAMPLING AND CROSS-ENTROPY

We briefly review some basics concepts and set up the notation. Let ℓ be the expected performance of a stochastic system given in the form

$$\ell(x) := P_f(\mathcal{M}(\boldsymbol{Y}) \ge x) = \mathbb{E}_f\left[I_{\{\mathcal{M}(\boldsymbol{Y})\ge x\}}\right], \tag{1}$$

where $\mathcal{M}(\boldsymbol{Y})$ is the *sample performance* and the subscript f means that the expectation of the random vector $\boldsymbol{Y}$ is taken with respect to the probability density function (pdf) f. Throughout this paper, the concept of "probability density function" should be understood in a broader way, that is, all the developments are valid when $\boldsymbol{Y}$ has a discrete distribution — in which case pdf's are replaced by probability mass functions (pmf's).

Let $G(\boldsymbol{y})$ be a probability measure (distribution) such that $\mathrm{d}G(\boldsymbol{y}) = g(\boldsymbol{y})\mathrm{d}\boldsymbol{y}$, where $g(\boldsymbol{y})$ is a pdf. Assume that $g(\boldsymbol{y})$ dominates $I_{\{\mathcal{M}(\boldsymbol{y})\ge x\}} f(\boldsymbol{y})$ in the absolutely continuous sense, that is, $\mathrm{supp}\{I_{\{\mathcal{M}(\boldsymbol{y})\ge x\}} f(\boldsymbol{y})\} \subset \mathrm{supp}\{g(\boldsymbol{y})\}$, where "supp" denotes the *support* of the corresponding function, i.e., the set of points where the function is not equal to zero. Using the pdf g we can represent $\ell(x)$ as

$$\ell(x) \;=\; \mathbb{E}_g\left[I_{\{\mathcal{M}(\boldsymbol{Z})\ge x\}}\,\frac{f(\boldsymbol{Z})}{g(\boldsymbol{Z})}\right].$$

An unbiased estimator of $\ell(x)$ is

$$\widehat{\ell}_N(x) \;=\; \frac{1}{N}\sum_{i=1}^{N} I_{\{\mathcal{M}(\boldsymbol{Z}_i)\ge x\}}\, W(\boldsymbol{Z}_i)\ , \tag{2}$$

where $W(z) = f(z)/g(z)$ is called the *likelihood ratio* (LR), and $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N$ are independent and identically distributed (i.i.d.) samples from $g(z)$.

The choice of the dominating pdf $g(\boldsymbol{y})$ is crucial for the variance of the LR estimator (2). Ideally, we would like to minimize the variance of $\widehat{\ell}_N$ with respect to the pdf g, that is, we want to solve

$$\min_g \ \mathrm{Var}_g\left[I_{\{\mathcal{M}(\boldsymbol{Z})\ge x\}}\frac{f(\boldsymbol{Z})}{g(\boldsymbol{Z})}\right]. \tag{3}$$

It is well known that the solution of problem (3) is

$$g^*(z) = \frac{I_{\{\mathcal{M}(z)\ge x\}} f(z)}{\int I_{\{\mathcal{M}(z)\ge x\}} f(z)\mathrm{d}z}. \tag{4}$$

The density $g^*(z)$ as per (4) is called the *optimal importance sampling density.* In general, however, implementation of the optimal importance sampling pdf $g^*(z)$ as per (4) is problematic. The main difficulty lies in the fact that in order to derive $g^*(z)$ one needs to know ℓ, which is the quantity we want to estimate from the simulation.

An alternative approach to the above problem can be derived when the underlying pdf's belong to some parametric family $\mathcal{F} = \{f(\boldsymbol{y}, \boldsymbol{v}),\ \ \boldsymbol{v} \in V\}$. *Throughout this paper, we will assume that this is the case.* Let $f(\boldsymbol{y}, \boldsymbol{u})$ denote the pdf of the random vector $\boldsymbol{Y}$ in (1). We then restrict the choice of the pdf g to pdf's from the same parametric family $\mathcal{F}$, so g differs from the original pdf $f(\boldsymbol{y}) = f(\boldsymbol{y}, \boldsymbol{u})$ by a single parameter (vector) $\boldsymbol{v}$. The likelihood ratio W in (2) with $g(\boldsymbol{y}) = f(\boldsymbol{y}, \boldsymbol{v})$ reduces to $W(\boldsymbol{Z}, \boldsymbol{u}, \boldsymbol{v}) = f(\boldsymbol{Z}, \boldsymbol{u})/f(\boldsymbol{Z}, \boldsymbol{v})$, where $\boldsymbol{v}$ $(\boldsymbol{v} \neq \boldsymbol{u})$ is called the *reference* parameter vector. It is readily seen that the optimal solutions of the variance-minimization problem (3) (with g restricted to $\mathcal{F}$) coincide with those of

$$\min_{\boldsymbol{v}\in V} \mathcal{V}(\boldsymbol{v}), \tag{5}$$

where

$$\mathcal{V}(\boldsymbol{v}) := \mathbb{E}_{\boldsymbol{v}_1}\left[I_{\{\mathcal{M}(\boldsymbol{X})\ge x\}} W(\boldsymbol{X}, \boldsymbol{u}, \boldsymbol{v}) W(\boldsymbol{X}, \boldsymbol{u}, \boldsymbol{v}_1)\right],$$

and $\boldsymbol{v}_1$ is chosen arbitrarily (to the extent that $f(z, \boldsymbol{v}_1)$ dominates $I_{\{\mathcal{M}(z)\ge x\}} f(z, \boldsymbol{u})$). This is a stochastic optimization problem, for which some methods such as *stochastic approximation* or *sample average approximation* (sometimes called stochastic counterpart) can be used — see, e.g., Rubinstein and Shapiro (1993). Lack of convexity, however, may lead to locally optimal solutions.

Another way to estimate the optimal reference parameter vector is based on the Kullback–Leibler *cross-entropy* (Kapur and Kesavan 1992), which defines a "distance" between the two probability distributions (densities) $f(\boldsymbol{y})$ and $g(\boldsymbol{y})$ and can be written as

$$\mathcal{D}(f, g) = \int f(\boldsymbol{y}) \ln \frac{f(\boldsymbol{y})}{g(\boldsymbol{y})}\, d\boldsymbol{y}. \tag{6}$$

Notice that $\mathcal{D}$ is not a distance in the formal sense, since in general $\mathcal{D}(f, g) \neq \mathcal{D}(g, f)$. Still, if $g = f$ then $\mathcal{D}(f, g) = 0$. A similar quantity can be defined for discrete distributions, with probability mass functions (pmf) in place of pds's and summations in place of integrals.

Let $\phi(z, \boldsymbol{u})$ denote the optimal measure in (4) with $f(z) = f(z, \boldsymbol{u})$. We can define a cross-entropy between $\phi(z, \boldsymbol{u})$ and $f(z, \boldsymbol{v})$, in analogy to (6), as

$$\mathcal{D}(\boldsymbol{v}) := \mathbb{E}_{\boldsymbol{u}}\left[\frac{I_{\{\mathcal{M}(\boldsymbol{Y})\ge x\}}}{c} \ln \frac{I_{\{\mathcal{M}(\boldsymbol{Y})\ge x\}} f(\boldsymbol{Y}, \boldsymbol{u})}{c f(\boldsymbol{Y}, \boldsymbol{v})}\right]$$

(where c is the denominator in (4)) and find the reference parameter vector $\boldsymbol{v}^*$ that solves $\min_{\boldsymbol{v}\in V}\ \mathcal{D}(\boldsymbol{v})$. It is obvious that the optimal solutions of this problem and of

$$\max_{\boldsymbol{v}\in V}\ D(\boldsymbol{v}), \tag{7}$$

where

$$D(\boldsymbol{v}) := \mathbb{E}_{\boldsymbol{v}_1}\left[I_{\{\mathcal{M}(\boldsymbol{X})\geq x\}}W(\boldsymbol{X},\boldsymbol{u},\boldsymbol{v}_1)\ln f(\boldsymbol{X},\boldsymbol{v})\right], \tag{8}$$

are identical. Given a sample $\boldsymbol{X}_1, ..., \boldsymbol{X}_N$ from $f(\boldsymbol{x},\boldsymbol{v}_1)$, we can estimate the optimal solution $\boldsymbol{v}^*$ of the above problem by solving

$$\max_{\boldsymbol{v}\in V}\ \widehat{D}_N(\boldsymbol{v}), \tag{9}$$

where $\widehat{D}_N(\boldsymbol{v}) = N^{-1}\sum_{i=1}^N I_{\{\mathcal{M}(\boldsymbol{X}_i)\geq x\}}W(\boldsymbol{X}_i,\boldsymbol{u},\boldsymbol{v}_1) \ln f(\boldsymbol{X}_i,\boldsymbol{v})$ is the sample average approximation of $D(\boldsymbol{v})$ in (8).

2.1 Relating Variance Minimization and Cross-Entropy

As seen above, both variance-minimization and the cross-entropy techniques (henceforth called VM and CE, respectively) have the same goal, namely, to approximate the optimal importance sampling density (4). The VM method ensures, by construction, the best approximation within the family $\{f(z,\boldsymbol{v}),\ \boldsymbol{v}\in V\}$ — in the sense that variance is minimized. The CE method, on the other hand, is based on a much nicer problem, which often has convexity properties and thus allows for computation of optimal solutions — sometimes even in closed form, see Section 3. Thus, it is natural to compare the solutions obtained from each method, in particular to check whether the easily computable CE-solution is close to the optimal VM-solution.

Consider the VM and CE problems in the form (5) and (7), respectively, with $\boldsymbol{v}_1 = \boldsymbol{u}$. It is clear that we can replace the objective function $D(\boldsymbol{v})$ in (7) by

$$D_1(\boldsymbol{v}) \ := \ -\mathbb{E}_{\boldsymbol{u}}\left[I_{\{\mathcal{M}(\boldsymbol{Y})\geq x\}}\ln W(\boldsymbol{Y},\boldsymbol{u},\boldsymbol{v})\right].$$

By noticing that $I^2 = I$ and conditioning on the event $\{\mathcal{M}(\boldsymbol{Y})\geq x\}$, we have

$$\begin{aligned}
\mathcal{V}(\boldsymbol{v}) &= \mathbb{E}_{\boldsymbol{u}}\left[I_{\{\mathcal{M}(\boldsymbol{Y})\geq x\}}W(\boldsymbol{Y},\boldsymbol{u},\boldsymbol{v})\right]\\
&= \mathbb{E}_{\boldsymbol{u}}\left[W(\boldsymbol{Y},\boldsymbol{u},\boldsymbol{v})\,|\,\mathcal{M}(\boldsymbol{Y})\geq x\right] \qquad (10)\\
&\quad \times P_{\boldsymbol{u}}(\mathcal{M}(\boldsymbol{Y})\geq x)\\
D_1(\boldsymbol{v}) &= -\,\mathbb{E}_{\boldsymbol{u}}\left[I_{\{\mathcal{M}(\boldsymbol{Y})\geq x\}}\ln W(\boldsymbol{Y},\boldsymbol{u},\boldsymbol{v})\right]\\
&= -\,\mathbb{E}_{\boldsymbol{u}}\left[\ln W(\boldsymbol{Y},\boldsymbol{u},\boldsymbol{v})\,|\,\mathcal{M}(\boldsymbol{Y})\geq x\right] \qquad (11)\\
&\quad \times P_{\boldsymbol{u}}(\mathcal{M}(\boldsymbol{Y})\geq x).
\end{aligned}$$

Notice the similarity between (10) and (11). Let now $\boldsymbol{v}^*$ be an optimal solution to the VM problem. Thus, we must have $\mathcal{V}(\boldsymbol{u},\boldsymbol{v}^*)-\mathcal{V}(\boldsymbol{u},\boldsymbol{v})\leq 0$ for all $\boldsymbol{v}\in V$, i.e.

$$\mathbb{E}_{\boldsymbol{u}}\left[W(\boldsymbol{Y},\boldsymbol{u},\boldsymbol{v})\ \frac{f(\boldsymbol{Y},\boldsymbol{v})-f(\boldsymbol{Y},\boldsymbol{v}^*)}{f(\boldsymbol{Y},\boldsymbol{v}^*)}\ \middle|\ \mathcal{M}(\boldsymbol{Y})\geq x\right]\ \leq\ 0 \tag{12}$$

for all $\boldsymbol{v}\in V$ .On the other hand, if $\boldsymbol{v}^*$ is an optimal solution to the CE problem then we must have

$$\mathbb{E}_{\boldsymbol{u}}\left[\ln\frac{f(\boldsymbol{Y},\boldsymbol{v})}{f(\boldsymbol{Y},\boldsymbol{v}^*)}\ \middle|\ \mathcal{M}(\boldsymbol{Y})\geq x\right]\ \leq\ 0 \quad \text{for all } \boldsymbol{v}\in V. \tag{13}$$

The solution sets defined by (12) and (13) are in general different. Suppose however that there exists $\boldsymbol{v}^*$ such that $f(\boldsymbol{y},\boldsymbol{v}^*)\geq f(\boldsymbol{y},\boldsymbol{v})$ for all $\boldsymbol{y}$ such that $\mathcal{M}(\boldsymbol{y})\geq x$ and all $\boldsymbol{v}\in V$. It is clear that such $\boldsymbol{v}^*$ satisfies both (12) and (13), i.e., such $\boldsymbol{v}^*$ is both VM- and CE-optimal. This suggests that, as x goes to infinity — i.e. as $P_{\boldsymbol{u}}(\mathcal{M}(Y)\geq x)$ goes to zero — the VM and CE problems tend to have the same solutions. The example in Section 4 corroborates that intuitive notion.

3 SPECIAL DISTRIBUTIONS

We discuss now ways to solve the CE problem (7). As it happens with (5), (7) is a stochastic optimization problem which can be solved by general techniques. It turns out, however, that for some families of distributions (7) can be solved *analytically*.

3.1 Natural Exponential Family

One important case occurs when the components of the random vector $\boldsymbol{Y}=(Y_1,\ldots,Y_n)$ are independent and each has a distribution in the *natural exponential family* (NEF) (see, e.g., Jorgensen 1997). A random variable X is said to have a NEF distribution if

$$f(y,w)=\exp(yw-k(w))h(y),\ \ w\in W\subset I\!R, \tag{14}$$

where $k(w)=\log\int e^{wy}h(y)dy$ is the cumulant function and $h(y)$ is a real valued (normalization) function of y. Many distributions, such as Poisson, exponential, etc., can be written as particular cases of the expression above; see Rubinstein and Melamed (1998) for details. It is possible to show that, if X has density $f(y,w)$ as in (14), then we have $\mu = \mathbb{E}X = k'(w)$ and $\mathrm{Var}[X] = k''(w)$. We then re-parameterize (14) as $\widetilde{f}(y,\mu) = \exp(yw(\mu) - k(w(\mu)))h(y)$, where $w(\mu) := [k']^{-1}(\mu)$ is the inverse function of k', which is well defined when k' is strictly increasing — which is the case if $\mathrm{Var}[X] > 0$. That is, the parameter of the distribution is its mean.

Because of the independence assumption and the ln function in (8), it is easy to see that problem (7) becomes separable. Moreover, by calculating the derivatives — which can done analytically — one can show that there is only one point $\boldsymbol{v}^* \in I\!R^n$ where the gradient of $D(\cdot)$ vanishes and, moreover, the Hessian matrix $\nabla^2 D(\boldsymbol{v}^*)$ is negative definite. It follows that v^* is the *unique global maximum* of $D(\cdot)$. By equating $\nabla D(\boldsymbol{v}^*)$ to zero, we obtain

$$v_j^* \; = \; \frac{\mathbb{E}_{\boldsymbol{u}}\left[Y_j I_{\{\mathcal{M}(\boldsymbol{Y})\geq x\}}\right]}{\mathbb{E}_{\boldsymbol{u}}\left[I_{\{\mathcal{M}(\boldsymbol{Y})\geq x\}}\right]}. \tag{15}$$

At first sight, formula (15) may seem useless since the denominator on the right hand side is the quantity $\ell(x)$ we want to estimate. Nevertheless, as we shall see later formula (15) is useful in terms of deriving an iterative algorithm. Also, note that for the variance minimization problem (5) there is no analytic solution similar to (15), even for NEF distributions. Thus, numerical optimization procedures must be used in such cases. This emphasizes one of the big advantages of the CE approach.

3.2 Finite Support Distributions

Another category of distributions for which the CE problem (7) can be conveniently solved is that of *finite support distributions*. Those distributions play an important role in rare event probability estimation, particularly due to their connection with combinatorial optimization problems; see Rubinstein (1999).

To proceed, suppose that the components of the random vector $\boldsymbol{Y} = (Y_1, \dots, Y_n)$ are independent. Assume $\boldsymbol{Y}_k \sim f(\boldsymbol{y}, \boldsymbol{u})$ takes on the values $y_{k1}, \dots, y_{km}$, and let $u_{kj} = P(Y_k = y_{kj})$. The goal is then to find a discrete distribution $f(\boldsymbol{y}, \boldsymbol{v})$ with independent marginals that solves the CE problem (7), where the set V is given by

$$V = \left\{ \boldsymbol{v} \in I\!R^{nm} : \sum_{j=1}^{m} v_{kj} = 1, \;\; k = 1, \dots, n, 0 \leq v_{kj} \leq 1 \right\}.$$

It easy to check that, in this case, (7) has concave objective function and linear constraints. Moreover, by the assumption of independence we have that f has a product form. It follows that the derivatives $\partial D / \partial v_{kj}$ are

$$\frac{\partial D}{\partial v_{kj}}(\boldsymbol{u}, \boldsymbol{v}) \; = \; \mathbb{E}_{\boldsymbol{u}^k}\left[I_{\{\mathcal{M}(\boldsymbol{Y})\geq x\}} \,|\, Y_k = y_{kj}\right] \frac{u_{kj}}{v_{kj}}.$$

In the above, $\mathbb{E}_{\boldsymbol{u}^k}$ denotes the expected value under $\boldsymbol{u}$ with respect to all components except Y_k (so $\boldsymbol{u}^k = (u_1, \dots, u_{k-1}, u_{k+1}, \dots, u_n)$). These derivatives, together with the sufficient Karush-Kuhn-Tucker optimality conditions for problem (7), yield an explicit solution, which can be expressed as

$$v_{kj}^* \; = \; \frac{\mathbb{E}_{\boldsymbol{u}^k}\left[I_{\{\mathcal{M}(\boldsymbol{Y})\geq x\}} \,\middle|\, Y_k = y_{kj}\right] u_{kj}}{\mathbb{E}_{\boldsymbol{u}}\left[I_{\{\mathcal{M}(\boldsymbol{Y})\geq x\}}\right]}, \tag{16}$$

provided of course that $P_{\boldsymbol{u}}(\mathcal{M}(\boldsymbol{Y}) \geq x) > 0$.

Finite support distributions also have the following important property:

Proposition 3.1 *Let x^* be the maximum value of $\mathcal{M}(\cdot)$ over the discrete set*

$$\mathcal{Y} = \{y_{11}, \dots, y_{1m}\} \times \dots \times \{y_{n1}, \dots, y_{nm}\},$$

and suppose that the maximizer of $\mathcal{M}(\cdot)$ over $\mathcal{Y}$ (call it $\boldsymbol{y}^*$) is unique. *Suppose that the random vector $\boldsymbol{Y}$ has independent components with discrete distribution on $\mathcal{Y}$. Then, the solution of both VM and CE programs (5) and (7) for $P(\mathcal{M}(\boldsymbol{Y}) \geq x^*)$ is the* atomic *measure (we shall also call it* degenerate*) with mass at $\boldsymbol{y}^*$.*

Proof. Let $\boldsymbol{v}_d^*$ denote the degenerate measure with mass on y^*. That $\boldsymbol{v}_d^*$ solves (5) follows immediately from the fact that the variance of estimator $\widehat{\ell}_N(x^*)$ given in (2), under $\boldsymbol{v}_d^*$, is *zero.*

Let $f(\boldsymbol{y}, \boldsymbol{u})$ denote the distribution of $\boldsymbol{Y}$. Consider now formula (16), derived for finite support distributions. Notice that the term $\mathbb{E}_{\boldsymbol{u}}\left[I_{\{\mathcal{M}(\boldsymbol{Y})\geq x^*\}} \,|\, Y_i = y_{ij}\right]$ is equal to zero if $y_{ij} \neq y_i^*$. Otherwise, we have

$$\begin{aligned} \mathbb{E}_{\boldsymbol{u}}\left[I_{\{\mathcal{M}(\boldsymbol{Y})\geq x^*\}} \,|\, Y_i = y_i^*\right] &= \mathbb{E}_{\boldsymbol{u}}\left[I_{\{\boldsymbol{Y}=y^*\}} \,|\, Y_i = y_i^*\right] \\ &= \prod_{k\neq i} P_{\boldsymbol{u}}(Y_k = y_k^*), \end{aligned}$$

and so in (16) we obtain that

$$v_{ij}^* \; = \; \begin{cases} 0 & \text{if } y_{ij} \neq y_i^* \\ \dfrac{\prod_{k\neq i} P_{\boldsymbol{u}}(Y_k = y_k^*) u_{ij}}{\prod_k P_{\boldsymbol{u}}(Y_k = y_k^*)} \; = \; 1 & \text{otherwise.} \end{cases}$$

■

Proposition 3.1 demonstrates the importance of finite support distributions — when x is the maximum value of $\mathcal{M}(\cdot)$, the solution of both VM and CE programs to estimate $P(\mathcal{M}(\boldsymbol{Y}) \geq x)$ are always the same, *regardless of the distribution of $\boldsymbol{Y}$*. This property in turn has nice implications for combinatorial optimization; see Rubinstein (1999) for a discussion.

It is also worth mentioning that the assumption of uniqueness of the maximizer of $\mathcal{M}$ in Proposition 3.1 can be artificially enforced by imposing some ordering on the finite set $\mathcal{Y}$, say the lexicographical order.

4 THE CE ALGORITHM

As mentioned before, formulas (15) and (16) are not intended for "stand-alone" use, as they depend on the quantity $\ell(x)$ we want to estimate. However, they do suggest a *multi-stage* procedure, which we describe now. The idea is to break down the "hard" problem of estimating the very small probability $\ell(x)$ into a sequence of "simple" problems, each time generating a sequence of pairs $\{(\widehat{\gamma_t}, \widehat{\boldsymbol{v}}_t)\}$ depending on the parameter (probability) ρ and such that $\rho >> \ell(x)$.

We start by choosing a not very small ρ, say $\rho = 10^{-2}$. Let γ_0 ($\gamma_0 < x$) be such that, under the original pdf $f(\boldsymbol{y}, \boldsymbol{u})$, the probability $\ell(\gamma_0) = \mathbb{E}_{\boldsymbol{u}}\left[I_{\{\mathcal{M}(\boldsymbol{Y}) \geq \gamma_0\}}\right]$ is at least ρ. We set next $\boldsymbol{v}_0 := \widehat{\boldsymbol{v}}_0 := \boldsymbol{u}$ and then proceed iterating in both $\boldsymbol{v}$ and γ with the goal of estimating the pair $\{\ell(x), \boldsymbol{v}^*\}$, as follows:

(a) **Adaptive estimation of** γ_t. For a fixed $\boldsymbol{v}_t$, let γ_t be a $(1-\rho)$-*quantile* of $\mathcal{M}(\boldsymbol{Z})$ under $\boldsymbol{v}_t$. That is, γ_t satisfies

$$P_{\boldsymbol{v}_t}(\mathcal{M}(\boldsymbol{Z}) \geq \gamma_t) \quad \geq \quad \rho, \tag{17}$$

$$P_{\boldsymbol{v}_t}(\mathcal{M}(\boldsymbol{Z}) \leq \gamma_t) \quad \geq \quad 1-\rho, \tag{18}$$

where $\boldsymbol{Z} \sim f(z, \boldsymbol{v}_t)$.

A simple estimate $\widehat{\gamma_t}$ of γ_t can be obtained by drawing a sample $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N$ from $f(z, \boldsymbol{v}_t)$ and taking the sample $(1-\rho)$-quantile. That is, we choose

$$\widehat{\gamma_t} = \widehat{\gamma_t}(\boldsymbol{v}_t) \equiv \mathcal{M}_{(t,\lceil(1-\rho)N\rceil)}, \tag{19}$$

where $\mathcal{M}_{(t,j)}$ is the j-th order statistics of the sequence $\mathcal{M}_{t,j} \equiv \mathcal{M}(\boldsymbol{Z}_{t,j}),\ \boldsymbol{Z}_{t,j} \equiv \boldsymbol{Z}_j,\ j = 1, \ldots, N$.

(b) **Adaptive estimation of** $\boldsymbol{v}_t$. For fixed γ_{t-1}, derive $\boldsymbol{v}_t$ from the solution of the program

$$\max_{\boldsymbol{v} \in V} \left\{\mathbb{E}_{\boldsymbol{v}_{t-1}}\left[I_{\{\mathcal{M}(\boldsymbol{Z}) \geq \gamma_{t-1}\}} W(\boldsymbol{Z}, \boldsymbol{u}, \boldsymbol{v}_{t-1}) \ln f(\boldsymbol{Z}, \boldsymbol{v})\right]\right\}. \tag{20}$$

The stochastic counterpart of (20) is as follows: for fixed $\widehat{\gamma}_{t-1}$, derive $\widehat{\boldsymbol{v}}_t$ from the following program

$$\max_{\boldsymbol{v} \in V} \left\{\frac{1}{N}\sum_{j=1}^{N} I_{\{\mathcal{M}(\boldsymbol{Z}_j) \geq \widehat{\gamma}_{t-1}\}} W(\boldsymbol{Z}_j, \boldsymbol{u}, \widehat{\boldsymbol{v}}_{t-1}) \ln f(\boldsymbol{Z}_j, \boldsymbol{v})\right\}. \tag{21}$$

As seen before, the optimal solutions of (20) and (21) can be obtained *analytically*, provided $f(\boldsymbol{y}, \boldsymbol{v})$ is either a NEF or a finite support distribution — cf. (15), (16). For example, the solution of (20) for NEF distributions is

$$v_{t,j} \;=\; \frac{\mathbb{E}_{\boldsymbol{v}_{t-1}}\left[Z_j I_{\{\mathcal{M}(\boldsymbol{Z}) \geq \gamma_{t-1}\}} W(\boldsymbol{Z}, \boldsymbol{u}, \boldsymbol{v}_{t-1})\right]}{\mathbb{E}_{\boldsymbol{v}_{t-1}}\left[I_{\{\mathcal{M}(\boldsymbol{Y}) \geq \gamma_{t-1}\}} W(\boldsymbol{Z}, \boldsymbol{u}, \boldsymbol{v}_{t-1})\right]}$$

whereas the solution of (21) is obtained by replacing expected values with sample averages in the above expression. Notice that, by construction, the above formula does not involve rare events.

Before presenting a complete description of the algorithm, let us discuss an example to illustrate the ideas. Suppose we are interested in estimating $\ell(x) = P(\mathcal{M}(\boldsymbol{Y}) \geq x)$, where $\mathcal{M}(\boldsymbol{Y}) = \min(Y_1, \ldots, Y_n)$ and the random variables $Y_1, \ldots, Y_n$ are exponentially identically distributed with mean u, i.e., $Y_i \sim f(y, u) = 1/u \exp(-y/u),\ i = 1, \ldots, n$. In this example, of course, we have $\ell(x) = e^{-nx/u}$, so there is no need for simulation. However, in order to illustrate the mechanism of the algorithm we will apply the multi-stage procedure described above. Moreover, the example motivates the need for some assumptions, which we will have to impose when dealing with presenting a complete formulation of the method.

Let us compute initially the CE-optimal parameter given by the solution to (7). As seen earlier, for the exponential distribution we can apply formula (15) directly. It follows that $v_i^* = v^* := x + u$. Notice that the VM-optimal solution to (5) is given by

$$v_i^* = \left[\frac{1}{u} + \frac{1}{x} - \sqrt{\frac{1}{u^2} + \frac{1}{x^2}}\right]^{-1}.$$

For $x >> u$, both methods yield $v^* \approx x$.

In order to measure the efficiency of the measure obtained, let us compute the *squared coefficient of variation* (SCV) of the LR estimator $\widehat{\ell_N}(x)$ in (2). This quantity, which also called *relative error,* gives an idea of how fast the sample size must grow in order to achieve a fixed precision; see, e.g., Rubinstein and Melamed (1998). It is given by

$$\kappa^2(\boldsymbol{v}, x) = \frac{N \mathrm{Var}[\widehat{\ell_N}(x)]}{\ell^2(x)}.$$

In the present example it is easy to see that

$$\kappa^2(v, x) = \left[\frac{v^2 e^{x/v}}{u(2v-u)}\right]^n - 1 \ .$$

For $v^* = x + u$, the above formula reduces to $\kappa^2(v^*, x) \approx x^n e^n/(2u)^n$. That is, for large x the SCV of the CE-optimal LR estimator increase in x *polynomially.* For $v = u$, which correspond to the crude Monte Carlo estimate, the SCV increase in x *exponentially*. For a more general discussion on complexity, see Asmussen and Rubinstein (1995).

Consider now γ_t defined in (17)-(18). Since the algorithm stops when $\gamma_t \geq x$, and since the distribution of $\mathcal{M}(\boldsymbol{Y})$ is continuous, we can write

$$\begin{aligned} \gamma_t &= \max\{\gamma \leq x \,:\, \exp(-\gamma n / v_t)) \geq \rho\} \\ &= \min\{x, C v_t/n\}, \end{aligned}$$

where $C = \log(1/\rho) > 0$. The parameter v_t defined in (20) can then be rewritten using (15) as

$$v_{t+1} \;=\; \gamma_t + u \;=\; \min\{x, C v_t/n\} + u. \qquad (22)$$

Consider the unidimensional function $g(v) = \min\{x, Cv/n\} + u$. It is easy to see that g has a single fixed point $\bar{v}$, and that $\bar{v} = v^* = x + u$ if and only if $x \leq (C/n)(x+u)$, i.e. $C \geq nx/(x+u)$. Since $C = \log(1/\rho)$, it follows that the CE procedure converges to the correct solution if and only if

$$\rho \;\leq\; \exp\left(-\frac{nx}{x+u}\right). \qquad (23)$$

Moreover, if $\rho \leq \exp(-n)$ (which implies (23)), i.e. if $C/n > 1$, then the differences $v_{t+1} - v_t$ *increase* until the point when x is hit by γ_t; otherwise, the differences $v_{t+1} - v_t$ *decrease* until the point when x is hit by γ_t.

At first sight, condition (23) seems discouraging, since it requires the parameter ρ to decrease exponentially with n. Notice however that this example constitutes an intrinsically difficult problem, since the probability being estimated goes to zero exponentially in n under *any* parameter. It makes perhaps more sense to consider the behavior of (23) for *fixed* n — then we see that $\rho \leq \exp(-n)$ is a sufficient condition for the CE algorithm to work, *regardless of the value of* x. We may also consider what happens when x is allowed to vary with n; for example, when $x = \Delta/n$ for some $\Delta > 0$, condition (23) becomes asymptotically $\rho \leq \exp(-\Delta/u)$.

The above example suggests that the value of the parameter ρ used in the CE algorithm plays a crucial role — as seen there, we can only expect the CE algorithm to converge to the correct values if ρ is sufficiently small. To determine a priori which ρ is acceptable, however, can be a difficult task. To overcome this problem, we can change the value of ρ *adaptively* (see Rubinstein 1999 for related ideas). Moreover, we shall also adopt an adaptive scheme to increase the sample size used in (19) and (21).

The complete algorithm is stated in detail below. It requires the definition of constants ρ (typically, $0.01 \leq \rho \leq 0.1$), $\alpha > 1$ and $\delta > 0$.

Algorithm 4.1:

1. *Set* $\rho_0 := \rho$, $N :=$ *initial sample size. Generate a sample* $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N$ *from the pdf* $f(\boldsymbol{z}, \boldsymbol{u})$ *and compute the sample* $(1-\rho_0)$*-quantile (19). Denote the initial solution by* $\widehat{\gamma}_0$*. Set t:=1.*
2. *Use the* **same** *current sample* $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N$ *to solve the stochastic program (21). Denote the solution by* $\widehat{\boldsymbol{v}}_t := \widehat{\boldsymbol{v}}_t(\widehat{\gamma}_{t-1})$.
3. *Generate a* **new** *sample* $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N$ *from the pdf* $f(\boldsymbol{z}, \widehat{\boldsymbol{v}}_t)$*. Let* $\rho_t := \rho$.
4. *Compute the sample* $(1-\rho_t)$*-quantile (19). Denote the solution by* $\widehat{\gamma}_t$.
5. *If* $\widehat{\gamma}_t \geq x$*, set* $\widehat{\gamma}_t := x$ *and solve the stochastic program (21) for* $\widehat{\gamma}_t = x$*. Denote the solution as* $\widehat{\boldsymbol{v}}_T$ *and go to step 7.*
6. *Otherwise, check whether there exists* $\bar{\rho}$ *such that the sample* $(1-\bar{\rho})$*-quantile of* $\mathcal{M}(\boldsymbol{Z}_1), \ldots, \mathcal{M}(\boldsymbol{Z}_N)$ *is bigger than or equal to* $\min\{x, \widehat{\gamma}_{t-1} + \delta\}$*:*
   - (a) *If there exists such* $\bar{\rho}$ *and* $\bar{\rho} = \rho_t$*, then set* $t := t+1$ *and reiterate from step 2;*
   - (b) *If there exists such* $\bar{\rho}$ *and* $\bar{\rho} < \rho_t$*, then set* $\rho_t := \bar{\rho}$ *and go back to step 4;*
   - (c) *Otherwise (i.e. if there exists no such* $\bar{\rho}$*) let* $N := \alpha N$ *and go back to step 3.*
7. *Estimate the rare-event probability* $\ell(x)$ *using the estimate (2), with* $\boldsymbol{v}_1$ *replaced by* $\widehat{\boldsymbol{v}}_T$*.*

5 CONVERGENCE OF THE CE METHOD

We discuss now some issues related to convergence of Algorithm 4.1. Let $\boldsymbol{v}^*$ be a CE-optimal solution, i.e. a maximizer of $D(\boldsymbol{v})$ defined in (7). That is, we have that

$$\boldsymbol{v}^* \in \operatorname{argmax}_{\boldsymbol{v} \in V} \left\{ \mathbb{E}_{\boldsymbol{u}} \left[I_{\{\mathcal{M}(\boldsymbol{Y}) \geq x\}} \ln f(\boldsymbol{Y}, \boldsymbol{v}) \right] \right\}. \qquad (24)$$

We will need the following assumption:

Assumption A: $P_{\boldsymbol{v}}(\mathcal{M}(\boldsymbol{Z}) \geq x) > 0$ for all $\boldsymbol{v} \in V$.

Assumption A simply ensures that the probability being estimated — $P_{\boldsymbol{u}}(\mathcal{M}(\boldsymbol{Z}) \geq x)$ — does not vanish when $\boldsymbol{u}$ is replaced by a feasible parameter $\boldsymbol{v} \in V$. The assumption is trivially satisfied when the distribution of $\mathcal{M}(\boldsymbol{Z})$ has infinite tail. For finite support distributions, the assumption holds as long as either x is less than the maximum value of the function $\mathcal{M}(\boldsymbol{Z})$, or if there is a positive probability that x is attained.

For $\boldsymbol{z} \in \mathbb{R}^n$, $\boldsymbol{v} \in \mathbb{R}^m$, and $\rho > 0$, define $\gamma(\boldsymbol{v}, \rho)$ as an arbitrary $(1-\rho)$-quantile of $\mathcal{M}(\boldsymbol{Z})$ under $\boldsymbol{v}$ (cf. (17)-(18)). Consider an arbitrary iteration t, and let $\rho_x^* := P_{\boldsymbol{v}_t}(\mathcal{M}(\boldsymbol{Z}) \geq x)$. By assumption A, $\rho_x^* > 0$. Let $\rho^* \in (0, \rho_x^*)$ be arbitrary.

By the definition of γ, we have that

$$\begin{aligned} P_{\boldsymbol{v}_t}\left(\mathcal{M}(\boldsymbol{Z}) \geq \gamma(\boldsymbol{v}_t, \rho^*)\right) &\geq \rho^* \\ P_{\boldsymbol{v}_t}\left(\mathcal{M}(\boldsymbol{Z}) \leq \gamma(\boldsymbol{v}_t, \rho^*)\right) &\geq 1 - \rho^* > 1 - \rho_x^*. \quad (25) \end{aligned}$$

Suppose that $\gamma(\boldsymbol{v}_t, \rho^*) < x$. Then,

$$P_{\boldsymbol{v}_t}\left(\mathcal{M}(\boldsymbol{Z}) \leq \gamma(\boldsymbol{v}_t, \rho^*)\right) \leq P_{\boldsymbol{v}_t}\left(\mathcal{M}(\boldsymbol{Z}) < x\right) = 1 - \rho_x^*,$$

which contradicts (25). It follows that $\gamma(\boldsymbol{v}_t, \rho) \geq x$ for ρ small enough and thus step 6 of Algorithm 4.1 can be accomplished provided $\widehat{\gamma}_t$ is also bigger than or equal to x. The proposition below shows that this happens for N large enough. In the proposition, the term "with probability one" refers to the probability space where $\boldsymbol{Z}$ lies, and when $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots$ are viewed as random variables on that space. We refer to Homem-de Mello and Rubinstein (2002) for a proof of this result.

Proposition 5.1 *Suppose assumption A holds. Let $\boldsymbol{v} \in V$, and let $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots$ be i.i.d. with common pdf $f(z, \boldsymbol{v})$. Let $\widehat{\gamma}_N(\boldsymbol{Z}, \rho)$ be a sample $(1-\rho)$-quantile of $\mathcal{M}(\boldsymbol{Z}_1), \ldots, \mathcal{M}(\boldsymbol{Z}_N)$. Then, there exists $\rho_x > 0$ and a random $N_x > 0$ such that, with probability one, $\widehat{\gamma}_N(\boldsymbol{Z}, \rho) \geq x$ for all $\rho \in (0, \rho_x)$ and all $N \geq N_x$. Moreover, the probability that $\widehat{\gamma}_N(\boldsymbol{Z}, \rho) \geq x$ for a given N goes to one* exponentially fast *with N.*

By the above result, at some iteration T we have $\widehat{\gamma}_T \geq x$ and thus in step 5 we set $\widehat{\gamma}_T := x$. It follows that we can view $\widehat{\boldsymbol{v}}_T$ as the solution of the problem

$$\max_{\boldsymbol{v} \in V} \left\{ \frac{1}{N} \sum_{j=1}^{N} I_{\{\mathcal{M}(\boldsymbol{Z}_j) \geq x\}} W(\boldsymbol{Z}_j, \boldsymbol{u}, \widehat{\boldsymbol{v}}_{T-1}) \ln f(\boldsymbol{Z}_j, \boldsymbol{v}) \right\}$$

which is precisely the sample average approximation problem (9). This is summarized in the following proposition.

Proposition 5.2 *Suppose that Assumption A holds. Then, Algorithm 4.1 converges w.p.1 to a solution of (9) after a finite number of iterations.*

We can then compare the approximating solution $\widehat{\boldsymbol{v}}_T$ and the "true" solution $\boldsymbol{v}^*$ using the asymptotic analysis for optimal solutions of stochastic optimization problems discussed in Rubinstein and Shapiro (1993). Following that approach, we obtain initially a *consistency* result: as $N \to \infty$, the distance between $\widehat{\boldsymbol{v}}_T$ and the solution set defined in (24) goes to zero (w.p.1) provided that: i) the function $\ln f(z, \boldsymbol{v})$ is continuous in $\boldsymbol{v}$, ii) the set V defined in assumption A is compact, and iii) there exists a function $h(z)$ such that $\mathbb{E}_{\boldsymbol{u}}[h(\boldsymbol{Z})] < \infty$ and $|\ln f(z, \boldsymbol{v})| \leq h(z)$ for all z and all $\boldsymbol{v} \in U$. Distributional results can also be obtained, see Rubinstein and Shapiro (1993).

Notice that the constant δ is used in Algorithm 4.1 only to ensure convergence. In practice, one can take $\delta = 0$ until the sequence $\{\widehat{\gamma}_t\}$ gets "stalled", at which point a positive δ is used again. Also, it is important to observe that, even if the optimal $\boldsymbol{v}^*$ could be obtained, some problems might still require a very large sample size in (2); see the discussion following the example in Section 4. Given the limitations of one's computational budget, Algorithm 4.1 can be used to detect such situation — the algorithm can be halted once ρ_t in step 6 gets too small (or, equivalently, when N gets too large).

6 NUMERICAL RESULTS

To illustrate the ideas set forth in the previous sections, we present now numerical results obtained for a manufacturing problem. In all examples below, we used an implementation of Algorithm 4.1 described in Section 5. Recall that the algorithm requires the definition of three constants ρ, α and δ. We used $\rho = 0.1$ and $\alpha = 2$. For δ, we adopted the conservative approach $\delta = 0$ (recall the discussion following the description of Algorithm 4.1). In these examples, such δ sufficed, i.e., the sequence $\{\widehat{\gamma}_t\}$ never got stalled. Moreover, step 6(c) of Algorithm 4.1 was never necessary, i.e. the initial sample size (determined after some pilot studies) was large enough.

Consider a single stage in a production system in which there are K single-server stations and a set of J jobs that must be processed sequentially by all stations in a prescribed order. We assume that the processing of job j on station k is a random variable whose distribution is known, and that each station processes its coming jobs on a first-come-first-serve basis, holding waiting jobs in a queue of infinite capacity. All jobs are released at time zero to be processed by the first station (this assumption is made just for notational convenience and can easily be dropped). For a job j, ($j = 1, ..., J$) and a station k, ($k = 1, ..., K$), let Y_{kj} denote the service time of processing job j on station k, and let C_{kj} denote the *completion time*, i.e., the time job j finishes its service at station k. By $\boldsymbol{Y} := (Y_{11}, \ldots, Y_{KJ})$ we denote the vector of service times, which is assumed to be random with a known distribution. Note that C_{Kj} can be viewed as a total completion time of job j and that each C_{kj} is a function of $\boldsymbol{Y}$, and hence is random. The above model is studied in Homem-de-Mello, Shapiro, and Spearman (1999) in the context of optimizing the performance system with respect to the release times of the jobs; we refer to that paper for details.

Our goal is to estimate the probability that all J jobs will be completed by a certain time x; that is, with $\mathcal{M}(\boldsymbol{Y}) = C_{KJ}(\boldsymbol{Y})$, we want to estimate $\ell(x) = P(\mathcal{M}(\boldsymbol{Y}) \geq x)$. Calculation of $\mathcal{M}(\boldsymbol{Y})$ for a particular realization of $\boldsymbol{Y}$ can be done via the recursive formula

$$C_{kj} = \max(C_{k-1,j},\ C_{k,j-1}) + Y_{kj}, \qquad (26)$$

with $C_{k0} = C_{0j} = 0$ for $k = 1, \ldots, K,\ j = 1, \ldots, J$. Notice that we can also view the above formula as a solution of a longest path problem in a directed graph; we refer again to Homem-de-Mello, Shapiro, and Spearman (1999) for details. Notice also that the above problem is *static* (which is the focus of the present paper) since the number of jobs under consideration is finite.

6.1 First Case: Exponential Distributions

We consider initially the case where all service times have exponential distribution. For simplicity, we assume that the service times of all jobs are i.i.d. with mean μ. In that case it is easy to check from (26) that

$$C_{KJ} \;\geq\; Y_{11} + \ldots + Y_{1J} + Y_{2J} + \ldots + Y_{KJ}.$$

The expression on the right hand side of the above inequality has distribution Gamma$(K + J - 1, \mu)$, so $P(\text{Gamma}(K + J - 1, \mu) \geq x)$ provides a lower bound on $P(\mathcal{M}(\boldsymbol{Y}) \geq x)$. To obtain an upper bound, we consider the Chebyshev inequality

$$P(\mathcal{M}(\boldsymbol{Y}) \geq x) \;\leq\; \frac{\mathbb{E}_\mu\left[\mathcal{M}(\boldsymbol{Y})^p\right]}{x^p}, \qquad (27)$$

which is valid for any $p > 0$ (note that $\mathcal{M}(\boldsymbol{Y}) \geq 0$ in this example). We consider three values for x, namely, $x = 0.8\Gamma$, $x = \Gamma$ and $x = 2\Gamma$, where $\Gamma = JK\mu$.

To estimate $P(\mathcal{M}(\boldsymbol{Y}) \geq x)$, we used the CE approach described in the previous sections. The parameter obtained — a $K \times J$-dimensional vector — determined the importance sampling distribution used to estimate the probability. For the sake of comparison, we also estimated the same probability using standard Monte Carlo. To provide a fair comparison, we provided the same *computational budget* for both methods. That is, we used a larger sample size for the crude Monte Carlo, since the CE methods requires extra computational time to calculate the optimal parameters. We increased the sample size sequentially until the total CPU time used by the crude Monte Carlo was the same as the time used for the CE method. The same stream of random numbers was used for the Monte Carlo and CE estimates for each x. The above procedure was replicated 100 times, and the average and a simultaneous 90% confidence interval were built from those 100 independent estimates, both for Monte Carlo and CE.

Table 1 displays the estimation results for $J = 10$, $K = 5$, $\mu = 25$, whereas Table 2 lists the lower and upper bounds. Although these results correspond to a particular instance of data, we must emphasize that similar type of results were observed for other problems we generated randomly. In the table, $\widehat{\ell}_N(x)$ is the estimate for $P(\mathcal{M}(\boldsymbol{Y}) \geq x)$, "90% H.W." denotes the half-width of a 96.67% confidence interval and N is the sample size. Notice that, since the sample size used with the Monte Carlo method was variable, the N column displays the average (as well as the half-width of a 96.67% confidence interval). Also, observe that the individual confidence of the three intervals displayed on the rows corresponding to each x is 96.67%; by Bonferroni's inequality, the *overall* confidence on those intervals is at least 90%.

Table 1: Estimated Probabilities for Exponential Distribution Case, $J = 10$, $K = 5$, $\mu = 25$

	MC		
x	$\widehat{\ell}_N(x)$	90% H.W.	N
1000	7.715×10^{-5}	2.719×10^{-5}	4751 (±117)
1250	0.000	0.000	12931 (±176)
2500	0.000	0.000	46430 (±786)

	CE		
x	$\widehat{\ell}_N(x)$	90% H.W.	N
1000	4.964×10^{-5}	1.044×10^{-5}	1000
1250	2.679×10^{-8}	9.687×10^{-9}	2000
2500	2.987×10^{-27}	2.903×10^{-27}	5000

Table 2: Estimated Bounds for Exponential Distribution Case, $J = 10$, $K = 5$, $\mu = 25$

x	lower bound	upper bound (95% H.W.)
1000	6.675×10^{-7}	1.148×10^{-3} (4.113×10^{-4})
1250	5.065×10^{-10}	6.853×10^{-7} (5.294×10^{-7})
2500	6.855×10^{-28}	5.263×10^{-25} (4.651×10^{-25})

6.2 Second Case: Discrete Distributions

We now consider the case where all service times have discrete distributions with finite support. As before, we assume for the sake of simplicity that the service times of all jobs are i.i.d., the common distribution being uniform on the set $\{10, 20, 30, 40\}$.

Notice that, because the random variables take on a finite number of values, the maximum possible completion time Ψ can be found by setting each random variable to its maximum value and solving a longest-path problem (cf. Homem-de-Mello, Shapiro, and Spearman 1999). In the current case, because all service times are i.i.d. the total completion time corresponds to a sum of $K + J - 1$ service times, so $\Psi = (K + J - 1) \times 40$. However, such procedure does not determine the probability of the maximum value, since there are multiple paths corresponding to it. A lower bound for the probability is $(1/4)^{K+J-1}$. We estimated $P(\mathcal{M}(\boldsymbol{Y}) \geq x)$ for two values of x, based on the value of the maximum completion time Ψ. We took $x = 0.9\Psi$ and $x = \Psi$ (obviously, $P(\mathcal{M}(\boldsymbol{Y}) > \Psi) = 0$).

To estimate $P(\mathcal{M}(\boldsymbol{Y}) \geq x)$, we used the CE approach described in the previous sections. Notice that in this case the parameter to be determined — the probabilities of each value of each service time — is a $K \times J \times m$-dimensional vector. As before, we also estimated the same probability using standard Monte Carlo, and provided the same *computational budget* for both methods. The same stream of random numbers was used for the Monte Carlo and CE estimates for each x. The above procedure was replicated 50 times, and the average and a simultaneous 90% confidence interval were built from those 50 independent estimates, both for Monte Carlo and CE.

Table 3 below displays the results for $J = 10$, $K = 5$. In this case, $\Psi = 560$. A lower bound for the probability $P(\mathcal{M}(\boldsymbol{Y}) \geq \Psi)$ is $(1/4)^{14} = 3.725 \times 10^{-9}$, while the upper bound computed from (27) with $p = 60$ is $8.430 \times 10^{-5} \pm 9.931 \times 10^{-6}$. Although the bounds in this case are a little bit loose, we must emphasize that, for problems where the probabilities $P(Y_{kj} = y_{kj})$ were randomly generated — in which case one can often calculate the probability $P(\mathcal{M}(\boldsymbol{Y}) \geq \Psi)$ exactly — the confidence intervals obtained from the CE method usually included the true value; we refer to Homem-de Mello and Rubinstein (2002) for details.

Table 3: Estimated Probabilities for Discrete Distribution Case, $J = 10$, $K = 5$, $m = 4$, Uniform Distribution

	MC		
x	$\widehat{\ell}_N(x)$	90% H.W.	N
500	1.423×10^{-2}	1.421×10^{-3}	680 (±40)
560	0.000	0.000	10192 (±19)

	CE		
x	$\widehat{\ell}_N(x)$	90% H.W.	N
500	1.424×10^{-2}	2.621×10^{-3}	100
560	7.004×10^{-7}	4.637×10^{-7}	700

The above results indicate high efficiency of the CE method for estimation rare-event probabilities, where the naive Monte Carlo method fails. For events that are not very rare, the CE method may still help in terms of providing estimates with smaller variance.

REFERENCES

Asmussen, S., and R. Y. Rubinstein. 1995. Complexity properties of steady-state rare-events simulation in queueing models. In *Advances in Queueing: Theory, Methods and Open Problems*, ed. J. Dshalalow, 429–462. CRC Press.

de Boer et al. 2001. A tutorial on the cross-entropy method. Manuscript, available at `<wwwhome.cs.utwente.nl/~ptdeboer/ce/tutorial.html>`.

de Boer, P. T., D. P. Kroese, and R. Y. Rubinstein. 2001. A fast cross-entropy method for estimating buffer overflows in queueing networks. Manuscript, Technion, Israel.

Garvels, M. J. J., and D. P. Kroese. 1998. A comparison of RESTART implementations. In *Proceedings of the 1998 Winter Simulation Conference*, ed. D. J. Medeiros, E. F. Watson, J. S. Carson, and M. S. Manivannan, 601–609: IEEE Press.

Glasserman et al. 1999. Multilevel splitting for estimating rare event probabilities. *Operations Research* 47 (4): 585–600.

Glynn, P. W., and D. L. Iglehart. 1989. Importance sampling for stochastic simulations. *Management Science* 35 (11): 1367–1392.

Görg, C. 1999. Simulating rare event details of ATM delay time distributions with RESTART/LRE. In *Proceedings of the RESIM Workshop*. University of Twente, The Netherlands.

Heidelberger, P. 1995. Fast simulation of rare events in queueing and reliability models. *ACM Transaction of Modeling and Computer Simulation* 5 (1): 43–85.

Homem-de Mello, T., and R. Y. Rubinstein. 2002. Rare event probability estimation for static models via cross-entropy and importance sampling. Manuscript, Ohio State University.

Homem-de-Mello, T., A. Shapiro, and M. L. Spearman. 1999. Finding optimal material release times using simulation based optimization. *Management Science* 45:86–102.

Jorgensen, B. 1997. *The theory of dispersion models*. Chapman and Hall.

Kapur, J. N., and H. K. Kesavan. 1992. *Entropy optimization principles with applications*. Academic Press.

Kovalenko, I. 1995. Approximations of queues via small parameter method. In *Advances in Queueing: Theory, Methods and Open Problems*, ed. J. Dshalalow, 481–509. CRC Press.

Rubinstein, R. Y. 1997. Optimization of computer simulation models with rare events. *European Journal of Operations Research* 99:89–112.

Rubinstein, R. Y. 1999. The cross-entropy method for combinatorial and continuous optimization. *Methodology and Computing in Applied Probability* 2:127–190.

Rubinstein, R. Y., and B. Melamed. 1998. *Modern simulation and modeling*. Chichester, England: J. Wiley & Sons.

Rubinstein, R. Y., and A. Shapiro. 1993. *Discrete event systems: Sensitivity analysis and stochastic optimization by the score function method*. Chichester, England: J. Wiley & Sons.

Shahabuddin, P. 1995. Rare event simulation of stochastic systems. In *Proceedings of the 1995 Winter Simulation Conference*, ed. C. Alexopoulos, K. Kang, W. R. Lilegdon, and D. Goldsman, 178–185: IEEE Press.

Villén-Altamirano, M., and J. Villén-Altamirano. 1999. About the efficiency of RESTART. In *Proceedings of the RESIM Workshop*, 99–128. University of Twente, The Netherlands.

AUTHOR BIOGRAPHIES

TITO HOMEM-DE-MELLO is an Assistant Professor in the Department of Industrial, Welding and Systems Engineering at the Ohio State University, Columbus, OH. He received his PhD degree in Operations Research from the Georgia Institute of Technology in 1998. His research interests include stochastic optimization, simulation analysis methodology, and stochastic models for inventory and revenue management. His e-mail and web addresses are `<homem-de-mello.1@osu.edu>` and `<www-iwse.eng.ohio-state.edu/isefaculty/tito/tito.htm>`.

REUVEN RUBINSTEIN holds a chair in Management Science at the Faculty of Industrial Engineering and Management of the Technion, Haifa, Israel, which he joined in 1973. Since then, he has visited many universities and research centers around the world, among them University of Illinois, Urbana, Harvard University, George Washington University, IBM Research Center, Bell Laboratories, Holmdel, NJ, NEC, and the Institute of Statistical Math., Japan. He is a member of several societies including the Operations Research Society of Israel and the American Operations Research Society. His e-mail and web addresses are `<ierrr01@ie.technion.ac.il>` and `<ie.technion.ac.il/ierrr01.phtml>`.