A Study on the Cross-Entropy Method for Rare-Event Probability Estimation

Tito Homem-de-Mello
Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois 60208, USA, tito@northwestern.edu

We discuss the problem of estimating probabilities of rare events in static simulation models using the recently proposed *cross-entropy method*, which is a type of importance-sampling technique in which the new distributions are successively calculated by minimizing the cross-entropy with respect to the ideal (but unattainable) zero-variance distribution. In our approach, by working on a functional space we are able to provide an efficient procedure without assuming any specific family of distributions. We then describe an implementable algorithm that incorporates the ideas described in the paper. Some convergence properties of the proposed method are established, and numerical experiments are presented to illustrate the efficacy of the algorithm.

Key words: cross entropy; importance sampling; rare events; simulation

History: Accepted by Susan M. Sanchez, Area Editor for Simulation; received April 2004; revised May 2005, August 2005, January 2006; accepted January 2006.

---

1. Introduction

A common problem in many areas of operations research is that of evaluating the expected value of a random quantity such as

$$\alpha \;:=\; \mathbb{E}_f\left[\mathcal{N}(Z)\right], \tag{1}$$

where $Z = (Z_1, \ldots, Z_n) \in I\!R^n$ is a vector with joint probability density function (pdf) $f(z)$, and $\mathcal{N}$ is an arbitrary real-valued function in $I\!R^n$. Many techniques have been developed over the years to provide estimates of α that are "better" than basic Monte Carlo, in the sense that the variance of the resulting estimates is reduced. See, for instance, Fishman (1997) and Law and Kelton (2000) for general discussions.

One method that has proven useful in many settings is the so-called *importance-sampling* (IS) technique. This well-known technique consists of drawing independent and identically

distributed (i.i.d.) samples $Z^1, \ldots, Z^N$ from an appropriately chosen pdf $g(\cdot)$, and estimating α by

$$\widehat{\alpha}_N(g) \;=\; \frac{1}{N}\sum_{j=1}^{N} \mathcal{N}(Z^j)\frac{f(Z^j)}{g(Z^j)}. \tag{2}$$

The pdf $g(\cdot)$ must dominate $\mathcal{N}(\cdot)f(\cdot)$ in the absolutely continuous sense. That is, $\text{Supp}[\mathcal{N}(\cdot)f(\cdot)] \subset \text{Supp}[g(\cdot)]$, where "Supp" denotes the *support* of the corresponding function, i.e., the set of points where the function is not equal to zero. The choice of g (henceforth called an *IS distribution*) is critical for the effectiveness of this approach, and in fact much of the research on importance sampling focuses on determining appropriate IS distributions. We remark that our assumption that the underlying distributions have pdfs is made only to simplify the exposition. The discussion in the paper can be extended to more general distributions. For example, for discrete distributions, the pdf should be understood as a probability mass function rather than a density in the strict sense. Of course, in that case integrals should be interpreted as summations.

A common approach encountered in the literature is to restrict the choice of IS distributions to some parametric family, say, $\{g(\cdot,\theta) : \theta \in \Theta\}$. Then, one attempts to find the "best" (in some sense) parameter θ^*. For example, θ^* can be the parameter that minimizes the *variance* of the estimator $\widehat{\alpha}_N(g(\cdot,\theta))$; see Rubinstein and Shapiro (1993) for a discussion and Vázquez-Abad and Dufresne (1998) for an application. Another example is that of Oh and Berger (1992), who assume a certain form for the "optimal" parameter and develop an adaptive procedure to estimate it.

Recently, Rubinstein (2002) introduced a method to calculate the parameter for the IS distribution in the context of rare events, which he called the *cross-entropy* (CE) method. The idea is to calculate the parameter θ^* such that $g(\cdot,\theta^*)$ minimizes the *Kullback-Leibler cross entropy* with respect to the zero-variance pdf g^* (defined in Section 2). In general, the calculation of θ^* requires solving a certain stochastic optimization problem, but in certain cases an explicit formula can be derived for θ^*. One example is when the underlying random vector has independent components and the family of distributions is the so-called natural exponential family, parameterized by the mean (de Boer et al. 2005, Homem-de-Mello and Rubinstein 2002).

In this paper we propose a more general view of the cross-entropy method. In our setting, we do not restrict the choice of IS distributions to some parametric family; our only constraint is that g have a *product form*. The rationale for this restriction is that

sampling from an arbitrary multivariate distribution is known to be a difficult task, so by imposing a product form on g we ensure that the components of the vector Z can be sampled independently, which then reduces the sampling problem to a unidimensional one. The cross-entropy optimization problem is solved on a functional space, and an explicit solution (henceforth called *CE-optimal* distribution) is provided. As we shall see later, the results obtained with such an approach generalize ones found in the literature.

We also study the relationship between the pdf given by the cross-entropy problem and the product-form pdf that minimizes the variance of the estimator $\widehat{\alpha}_N(g)$. Our discussion suggests that the cross-entropy problem can be viewed as a slight variation of variance minimization, with the advantage that the underlying optimization problem can be solved in closed form. We then discuss an adaptive scheme to estimate the CE-optimal distribution. A basic version of the procedure — for parametric distributions — was proposed by Rubinstein (1999); here, we propose an algorithm that allows more general distributions and incorporates automatic adjustment of the parameters of the algorithm. A detailed comparison between our approach and the standard CE method is provided in Section 3.2. Some aspects related to convergence of the proposed algorithm are discussed in Section 3.3, where we establish that the adaptive procedure finishes after finitely many iterations. The estimate obtained can then be refined by increasing the sample size if necessary.

Finally, we present some numerical results in Section 4 for a flow-line production system, where the goal is to estimate the probability that a certain sequence of jobs finishes processing after a certain time. This is a difficult problem for which no general analytical solution is available. The results suggest that the CE method works very well, providing accurate estimates for probabilities as low as 10^{-56} in reasonable time. These results are checked through the derivation of lower and upper bounds, or even exact values in some cases. We also provide a numerical comparison with the hazard-rate twisting method described in Huang and Shahabuddin (2004) and Juneja and Shahabuddin (2002).

In summary, the main contributions of this paper are the following. (i) We provide a general framework for the cross-entropy method, which allows for derivation of closed-form solutions to the CE-optimization problem for arbitrary distributions. This generalizes prior work and consequently opens the way for use of the CE method with other types of distributions. (ii) We propose and test a modified version of the CE algorithm that incorporates the generalizations mentioned in (i) and *provably* finishes after finitely many iterations. We also illustrate the numerical behavior of the method on a difficult 50-dimensional problem

with large variability. (iii) Along the way, we provide a new result on convergence of quantile estimates when quantiles are not unique.

1.1. Literature Review

The basic ideas of importance-sampling were outlined by Kahn and Marshall (1953). Glynn and Iglehart (1989) extended these ideas to stochastic processes. Since then, a considerable amount of research has been devoted to the study of IS techniques in simulation, in particular for rare-event simulation; see Heidelberger (1995) and Shahabuddin (1995) for reviews. Most of the work in this area, however, deals with *dynamic* models, in the sense that α in (1) is calculated either from some steady-state performance measure or from some stopping time (e.g., the probability that a buffer exceeds a certain capacity). Among other techniques that have been proposed for dynamic problems are the *splitting* and *RESTART* methods — see, for instance, Glasserman et al. (1999) and Villén-Altamirano and Villén-Altamirano (1999). Applications of the CE method to dynamic systems are discussed in de Boer (2000) and Kroese and Rubinstein (2004).

In contrast, our model is essentially *static*, i.e., we want to estimate (1) for a given deterministic function $\mathcal{N}$ of a random vector Z of known distribution. Such a problem is encountered more often in the statistics literature, and in fact IS techniques have been studied in that context as well; some close references to our work are Oh and Berger (1992, 1993) and Zhang (1996). Huang and Shahabuddin (2004) discuss a general approach based on the hazard-rate twisting method of Juneja and Shahabuddin (2002) to estimate rare-event probabilities in static problems. That method, which is also used by Juneja et al. (2004) to estimate rare-event probabilities in stochastic PERT networks, is discussed in more detail in Section 4. Static problems have also gained importance in the simulation community because of the applications of these models in finance (see, e.g., Glasserman 2004). In addition, static rare-event problems have an interesting connection with combinatorial optimization (de Boer et al. 2005). We must mention that *large-deviations* techniques are often employed to estimate rare-event probabilities (e.g., Bucklew 1990, Dembo and Zeitouni 1998); however, these techniques are usually more applicable when the underlying quantities involve a sum of random variables, which is not necessarily the case of our setting as we deal with general functions $\mathcal{N}$.

2. Determining a Good IS Distribution

2.1. The Cross-Entropy Approach

It is well known that if $\mathcal{N}(\cdot) \geq 0$ in (1) then the optimal choice for the IS distribution is given by

$$g^*(z) = \frac{\mathcal{N}(z) f(z)}{\alpha}; \tag{3}$$

this yields a *zero-variance* estimator, i.e., $\mathrm{Var}[\widehat{\alpha}_N(g^*)] = 0$ in (2). Of course, we cannot use g^* since it depends on the quantity α we want to estimate. However, even if we could somehow compute g^*, it would be difficult to generate samples from it, since g^* is a joint pdf. One way to circumvent the latter issue is to determine the distribution $\bar{g}$ that minimizes the Kullback-Leibler "distance" to g^* among all densities with a product form, i.e. $g(z) = g_1(z_1) \times \cdots \times g_n(z_n)$. The Kullback–Leibler cross-entropy (see Kullback and Leibler 1951, Kapur and Kesavan 1992) defines a "distance" between two pdf's $f(\cdot)$ and $g(\cdot)$ as

$$\mathcal{D}(f, g) = \int_{I\!R^n} f(z) \log \frac{f(z)}{g(z)}\, dz.$$

Notice that $\mathcal{D}$ is not a distance in the formal sense, since in general $\mathcal{D}(f, g) \neq \mathcal{D}(g, f)$. Nevertheless, it is possible to show (Kullback and Leibler 1951) that $\mathcal{D}(f, g) \geq 0$ and that $\mathcal{D}(f, g) = 0$ if and only if the corresponding cdfs are the same. The problem can then be defined as

$$\min \{\mathcal{D}(g^*, g) \,:\, g \in \mathcal{G}\}, \tag{4}$$

where $\mathcal{G}$ is the set of densities with product form such that $\mathrm{Supp}[\mathcal{N}(\cdot) f(\cdot)] \subset \mathrm{Supp}[g(\cdot)]$. For an arbitrary $g \in \mathcal{G}$ we have

$$\begin{aligned} \mathcal{D}(g^*, g) &= \int_T g^*(z) \log \left[\frac{g^*(z)}{g(z)} \right] dz \\ &= \int_T g^*(z) \log[g^*(z)]\, dz - \frac{1}{\alpha} \int_T \mathcal{N}(z) f(z) \log[g(z)]\, dz \end{aligned}$$

where $T \subset I\!R^n$ denotes the support of $\mathcal{N}(\cdot) f(\cdot)$. We will use the convention that $a \log[0] = -\infty$ if $a > 0$. It follows that the solution of (4) is the same as the solution of the problem

$$\min_{g \in \mathcal{G}} \; -\int_T \mathcal{N}(z) f(z) \log[g(z)]\, dz \;=\; \min_{g \in \mathcal{G}} \; -\mathbb{E}_f \left[\mathcal{N}(Z) \log[g(Z)] \right]. \tag{5}$$

Let $f_{Z_i}(\cdot)$ denote the marginal distribution of Z_i, and $\tilde{Z}$ denote the vector $(Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_n)$. Define the function

$$\varphi_i(z_i) \;:=\; \mathbb{E}_{\tilde{Z}|Z_i} \left[\mathcal{N}(Z) \,|\, Z_i = z_i \right] \, f_{Z_i}(z_i), \tag{6}$$

where $\mathbb{E}_{\tilde{Z}|Z_i}[\cdot]$ denotes the expectation with respect to the conditional density $f_{\tilde{Z}|Z_i}(\cdot)$ of $\tilde{Z}$ given Z_i. Notice that we can write $\alpha = \mathbb{E}_f[\mathcal{N}(Z)] = \int_T \mathcal{N}(z)f(z)\,dz = \int_{T_i} \varphi_i(z_i)\,dz_i$, where $T_i \subset I\!R$ is the support of $\varphi_i(\cdot)$. Moreover, since $\mathbb{E}[\log[g_i(Z_i)]\mathcal{N}(Z)] = \mathbb{E}[\mathbb{E}[\log[g_i(Z_i)]\mathcal{N}(Z) \mid Z_i]] = \mathbb{E}[\log[g_i(Z_i)]\mathbb{E}[\mathcal{N}(Z) \mid Z_i]]$, we have that $\int_T \log[g_i(z_i)]\mathcal{N}(z)f(z)\,dz = \int_{T_i} \log[g_i(z_i)]\varphi_i(z_i)\,dz_i$ and thus minimizing $\mathcal{D}(g^*, g)$ subject to $g \in \mathcal{G}$ is equivalent to solving the functional problem

$$\begin{array}{rcl} \max\limits_{g_i \in Q} & & \displaystyle\int_{T_i} \log[g_i(z_i)]\varphi_i(z_i)\,dz_i \\ \text{s.to} & & \\ \displaystyle\int_{T_i} g_i(z_i)\,dz_i & = & 1. \end{array} \tag{7}$$

In (7), Q is the subset of L^1 (integrable functions) consisting of nonnegative functions whose support is T_i. Notice that Q is a convex set. Moreover, Q is non-empty since it contains $\varphi_i(\cdot)$.

We now discuss ways to solve (7). Define the functionals (in L^1)

$$\begin{array}{rcl} F_i(g_i) & = & \displaystyle\int_{T_i} \log[g_i(z_i)]\varphi_i(z_i)\,dz_i \\ H_i(g_i) & = & \displaystyle\int_{T_i} g_i(z_i)\,dz_i - 1. \end{array}$$

It is clear that F_i is *concave* on Q, whereas H_i is *affine* on L^1. Let us compute the derivatives of these functionals, which we denote respectively by $DF_i(g_i)$ and $DH_i(g_i)$. These derivatives are operators in L^1, defined as

$$\begin{array}{rcl} DF_i(g_i)h & = & \displaystyle\lim_{t\to 0} \frac{F_i(g_i + th) - F_i(g)}{t} \\ & = & \displaystyle\lim_{t\to 0} \int_{T_i} \frac{\log[g_i(z_i) + th(z_i)] - \log[g_i(z_i)]}{t}\,\varphi_i(z_i)\,dz_i. \end{array} \tag{8}$$

Since $\log[\cdot]$ is concave, the function $\rho(t) := (\log[x + td] - \log[x])/t$ is monotone in t for any $x > 0$ and any d. It follows from the monotone convergence theorem that we can interchange the integral and the limit in (8) and hence we obtain

$$DF_i(g_i)h = \int_{T_i} \frac{h(z_i)}{g_i(z_i)}\,\varphi_i(z_i)\,dz_i.$$

Similarly, we have $DH_i(g_i)h = \int_{T_i} h(z_i)\,dz_i$.

Consider now the Lagrangian functional associated with (7), which is defined on $L^1 \times I\!R$ as $L_i(g_i, \lambda) := F_i(g_i) + \lambda H_i(g_i)$. It is known (Bonnans and Shapiro 2000, Proposition 3.3)

that if there exists a pair $(\bar{g}_i, \bar{\lambda})$ such that $\bar{g}_i \in Q$, $H_i(\bar{g}_i) = 0$ and

$$\bar{g}_i \;\in\; \underset{g_i \in Q}{\operatorname{argmax}}\; L_i(g_i, \bar{\lambda}), \tag{9}$$

then $\bar{g}_i$ solves (7). The proposition below exhibits such a pair:

Proposition 1 *Consider the function $\varphi_i(\cdot)$ defined in* (6)*, and define the density function*

$$\bar{g}_i(z_i) \;:=\; \frac{\varphi_i(z_i)}{\alpha}. \tag{10}$$

Then, the pair $(\bar{g}_i, -\alpha)$ satisfies (9) and therefore $\bar{g}_i$ solves (7).

Proof. It is immediate that $\bar{g}_i \in Q$ and $H_i(\bar{g}_i) = 0$. Thus, we just need to check (9). From the definition of the Lagrangian function, we have that, for given g_i and λ,

$$\begin{aligned} DL_i(g_i, \lambda)h &= \int_{T_i} \frac{h(z_i)}{g_i(z_i)}\, \varphi_i(z_i)\, dz_i + \lambda \int_{T_i} h(z_i)\, dz_i \\ &= \int_{T_i} \left[\frac{\varphi_i(z_i)}{g_i(z_i)} + \lambda \right] h(z_i)\, dz_i. \end{aligned} \tag{11}$$

Consider now the function $\bar{g}_i$ defined in (10). It is clear that $\varphi_i(z_i)/g_i(z_i) - \alpha = 0$ for all z_i and thus from (11) we have that

$$DL_i(\bar{g}_i, -\alpha) \;\equiv\; 0. \tag{12}$$

Since the function $L_i(\cdot, \lambda)$ is concave on Q for any λ, (12) implies that $\bar{g}_i$ maximizes $L_i(\cdot, -\alpha)$. This concludes the proof. ■

Corollary 1 *Let $b(\cdot)$ be an arbitrary function. Then, the expected value of a random variable X_i with the density $\bar{g}_i$ defined in* (10) *is*

$$\mathbb{E}_{\bar{g}_i}\left[b(X_i)\right] \;=\; \frac{\mathbb{E}_f\left[b(Z_i)\mathcal{N}(Z)\right]}{\mathbb{E}_f\left[\mathcal{N}(Z)\right]}. \tag{13}$$

In particular, by taking $b(y) = y^k$ we obtain an expression for the kth moment of $\bar{g}_i$.

Proof. We have

$$\begin{aligned} \mathbb{E}_{\bar{g}_i}\left[b(X_i)\right] &= \int_{T_i} b(x_i)\, \bar{g}_i(x_i)\, dx_i \;=\; \frac{1}{\alpha} \int_{T_i} b(x_i) \mathbb{E}_{\tilde{Z}|Z_i}\left[\mathcal{N}(Z) \mid Z_i = x_i\right]\, f_{Z_i}(x_i)\, dx_i \\ &= \frac{1}{\alpha} \int_{T_i} \mathbb{E}_{\tilde{Z}|Z_i}\left[b(Z_i)\mathcal{N}(Z) \mid Z_i = x_i\right]\, f_{Z_i}(x_i)\, dx_i \\ &= \frac{\mathbb{E}_f\left[b(Z_i)\mathcal{N}(Z)\right]}{\alpha}. \blacksquare \end{aligned} \tag{14}$$

Proposition 1 and Corollary 1 generalize previous results. The exact form of the solution of the cross-entropy problem (4) had been derived only for particular cases — namely, parametric problems where the family of distributions is the natural exponential family parameterized by the mean (de Boer et al. 2005, Homem-de-Mello and Rubinstein 2002). Proposition 1 and Corollary 1, in turn, do not assume any distribution.

Proposition 1 gives the desired product-form distribution. While that facilitates the task of generating random samples from that distribution — since each component can be generated independently — we still must deal with the fact that $\bar{g}_i$ depends on α, the quantity we want to estimate. We address this in Section 3.

2.2. Relating Variance Minimization and Cross Entropy

In Section 2.1 we showed how the optimization problem of minimizing the Kullback-Leibler "distance" to the optimal distribution can be solved analytically. While such a property is certainly appealing, it is natural to inquire what type of properties the resulting pdf has.

We address this issue by comparing the cross-entropy problem (4) with the problem of finding a pdf that yields an estimate with minimum variance. Suppose we want to find a pdf g° with product form such that g° minimizes $\mathrm{Var}[\widehat{\alpha}_N(g)]$. Notice that

$$\mathrm{Var}_g\left[\mathcal{N}(Z)\frac{f(Z)}{g(Z)}\right] \;=\; \mathbb{E}_g\left[\left(\mathcal{N}(Z)\frac{f(Z)}{g(Z)}\right)^2\right]-\left(\mathbb{E}_g\left[\mathcal{N}(Z)\frac{f(Z)}{g(Z)}\right]\right)^2 \;=\; \mathbb{E}_f\left[\mathcal{N}(Z)^2\frac{f(Z)}{g(Z)}\right]-\alpha^2,$$

so minimizing the variance is equivalent to solving the problem

$$\min_{g\in\mathcal{G}} \quad \mathbb{E}_f\left[\mathcal{N}(Z)^2\tfrac{f(Z)}{g(Z)}\right] \quad . \tag{15}$$

In turn, (5) has the same solution as

$$\min_{g\in\mathcal{G}} \quad \mathbb{E}_f\left[\mathcal{N}(Z)\log\tfrac{f(Z)}{g(Z)}\right] \quad . \tag{16}$$

Notice the similarity between (15) and (16).

Consider now the particular case where $\mathcal{N}$ is an indicator function of the form $I_{\{\mathcal{M}(Z)\geq a\}}$ — which is the setting of this paper from Section 3 on. Then, by noticing that $I^2 = I$ and conditioning on the event $\{\mathcal{M}(Z) \geq a\}$, we have the solutions of (15) and (16) are respectively the same as the solutions of

$$\min_{g\in\mathcal{G}} \; \mathbb{E}_f\left[\frac{f(Z)}{g(Z)}\,|\,\mathcal{M}(Z)\geq a\right] \quad \text{and} \quad \min_{g\in\mathcal{G}} \; \mathbb{E}_f\left[\log\frac{f(Z)}{g(Z)}\,|\,\mathcal{M}(Z)\geq a\right].$$

Since log is an increasing function, we see that the two problems are indeed similar. Clearly, without the constraint $g \in \mathcal{G}$ the solution of both problems is the zero-variance pdf g^*.

3. Estimating Rare-Event Probabilities

We turn now to the issue of using the product-form distribution $\bar{g}_i$ derived in Proposition 1 to obtain estimates of the value α defined in (1) when α is the probability of a *rare event*. That is, the function $\mathcal{N}$ in (1) is of the form $\mathcal{N}(Z) = I_{\{\mathcal{M}(Z)\geq a\}}$ for some function $\mathcal{M}$ and some $a \in I\!R$ such that $\{\mathcal{M}(Z) \geq a\}$ is an event of small probability. In what follows, we describe an implementable algorithm and discuss some issues related to convergence.

3.1. The Algorithm

As remarked earlier, using $\bar{g}_i$ directly is impossible since it depends on α. To overcome this difficulty, we describe now a *multi-stage* algorithm for estimating $\bar{g}$, whose basic version was first proposed by Rubinstein (1999). The improvements we propose here include closed-form expressions (derived from the generalized approach of Section 2) and an automatic update of the main parameters of the algorithm.

The idea of the algorithm is to generate an increasing sequence of values $\{\widehat{\gamma}^k\}_{k=1,2,\ldots}$ and a sequence of distributions $\{\widehat{g}^k\}_{k=1,2,\ldots}$ such that $\widehat{g}^k$ is a good importance-sampling distribution to estimate $P_f(\mathcal{M}(Z) \geq \widehat{\gamma}^k)$. This is accomplished by solving the cross-entropy problem (4) with the underlying function $\mathcal{N}(Z)$ set to $I_{\{\mathcal{M}(Z)\geq\widehat{\gamma}^k\}}$. Notice however that the solution to (4), which is given by (10), depends on the quantity $P_f(\mathcal{M}(Z) \geq \widehat{\gamma}^k)$. The latter expression can be written as

$$P_f(\mathcal{M}(Z) \geq \widehat{\gamma}^k) \;=\; \mathbb{E}_{\widehat{g}^{k-1}}\left[I_{\{\mathcal{M}(Z)\geq\widehat{\gamma}^k\}}\frac{f(Z)}{\widehat{g}^{k-1}(Z)}\right] \tag{17}$$

provided the condition $f(z) > 0 \Longrightarrow \widehat{g}^{k-1}(z) > 0$ holds for all z such that $\mathcal{M}(z) \geq \widehat{\gamma}^k$. By construction, $\widehat{g}^{k-1}$ is a good distribution to estimate $P_f(\mathcal{M}(Z) \geq \widehat{\gamma}^{k-1})$; thus, if $\widehat{\gamma}^k$ is not much bigger than $\widehat{\gamma}^{k-1}$ — i.e. if the event $\{\mathcal{M}(Z) \geq \widehat{\gamma}^k\}$ is not rare under $\widehat{g}^{k-1}$ — one expects $\widehat{g}^{k-1}$ to be a reasonably good distribution to estimate $P_f(\mathcal{M}(Z) \geq \widehat{\gamma}^k)$ as well. Once $\widehat{\gamma}^k$ reaches the threshold value a, then the algorithm returns the current density $\widehat{g}^k$.

We provide now a formal description of the algorithm. Let $\gamma(g, \rho)$ denote an arbitrary $(1-\rho)$-quantile of $\mathcal{M}(Z)$ under g, i.e., $\gamma(g, \rho)$ satisfies

$$P_g(\mathcal{M}(Z) \geq \gamma(g,\rho)) \;\geq\; \rho, \tag{18}$$

$$P_g(\mathcal{M}(Z) \leq \gamma(g,\rho)) \;\geq\; 1-\rho. \tag{19}$$

Notice that, given an i.i.d. sample $Z^1, \ldots, Z^N$ from $g(\cdot)$, $\gamma(g, \rho)$ can be easily estimated by a $(1-\rho)$-*sample* quantile of $\mathcal{M}(Z^1), \ldots, \mathcal{M}(Z^N)$. The latter quantity is denoted by $\widehat{\gamma}_{_N}(Z, \rho)$.

The algorithm requires the definition of constants ρ^0 (typically, $0.01 \leq \rho^0 \leq 0.1$), $\nu > 1$ and $\delta > 0$. Below, an expression of the form $\Theta(Z^j)|Z_i^j = z_i$ denotes $\Theta(Z_1^j, \ldots, Z_{i-1}^j, z_i, Z_{i+1}^j, \ldots, Z_n^j)$.

Algorithm 1 :

1. *Set $k := 1$, $N :=$ initial sample size, $\widehat{g}^0 := f$.*

2. *Generate i.i.d. samples $Z^1, \ldots, Z^N$ from the pdf $\widehat{g}^{k-1}(\cdot)$.*

3. *Let $\widehat{\gamma}^k := \min\{a, \widehat{\gamma}_{_N}(Z, \rho^{k-1})\}$.*

4. *Define*
$$\widehat{\alpha}^k \;:=\; \frac{1}{N}\sum_{j=1}^{N} I_{\{\mathcal{M}(Z^j)\geq\widehat{\gamma}^k\}}\frac{f(Z^j)}{\widehat{g}^{k-1}(Z^j)}.$$

5. *Compute the unidimensional density $\widehat{g}_i^k(\cdot)$ as*
$$\widehat{g}_i^k(z_i) \;:=\; \frac{\frac{1}{N}\sum_{j=1}^{N}\left(I_{\{\mathcal{M}(Z^j)\geq\widehat{\gamma}^k\}}\frac{f(Z^j)}{\widehat{g}^{k-1}(Z^j)}\;\middle|\; Z_i^j = z_i\right) f_{Z_i}(z_i)}{\widehat{\alpha}^k}.$$

6. *If $\widehat{\gamma}^k = a$, STOP; let $\widetilde{g} := \widehat{g}^k$ be the distribution returned by the algorithm.*

7. *Otherwise, let $C(\rho)$ denote the condition*

$$C(\rho) \;:\; \text{the sample } (1-\rho)\text{-quantile of } \mathcal{M}(Z^1), \ldots, \mathcal{M}(Z^N) \text{ is bigger than or equal to } \min\{a, \widehat{\gamma}^{k-1} + \delta\}$$

   (a) *If $C(\rho)$ is satisfied with $\rho = \rho^{k-1}$, then set $\rho^k := \rho^{k-1}$, $k := k+1$ and reiterate from step 2;*

   (b) *If $C(\rho)$ is not satisfied with $\rho = \rho^{k-1}$ but it is satisfied with some $\rho < \rho^{k-1}$, then set $\rho^k :=$ largest of such ρ and go back to step 3;*

   (c) *If $C(\rho)$ is not satisfied with any $\rho \leq \rho^{k-1}$, then let $N := \nu N$ and go back to step 2.*

Note that when the original density f is not of product form, extra care should be taken to use the samples $Z^1, \ldots, Z^N$ the first time step 5 is executed — after all, by fixing $Z_i^j = z_i$ the distribution of the other Z_k ($k \neq i$) change. One way around this problem is to generate new samples $Z^1, \ldots, Z^N$ for each value of z_i, using the same stream of random numbers for

all of them. From the second iteration on this is no longer necessary, since by construction $\widehat{g}^k(z)$ $(k \geq 1)$ has product form.

Usually we cannot compute the whole density function in step 5, as this would require infinitely many calculations. One case in which this can be easily accomplished is when Z has a *discrete* distribution with a small support. For example, suppose that the Z_i's are independent and each Z_i takes on values $\{z_{i1}, \ldots, z_{im}\}$; then, $f_{Z_i}(\cdot)$ is the probability mass function $f_{Z_i}(z_{ij}) = P(Z_i = z_{ij})$, so computation of $\widehat{g}^k$ is achieved by doing the calculation in step 5 for mn values.

In the case of continuous distributions, two possible approaches are: (i) to approximate the density by a discrete distribution, and (ii) to fix a family for the IS distributions and then compute some of its moments. For example, suppose we fix the family of normal distributions; then it suffices to compute the first two moments in order to specify $\widehat{g}^k$. The gamma distribution also shares that property. In that case, step 5 is replaced by the following step, derived from (14):

5′. Estimate the rth moment of $\widehat{g}_i^k$ by

$$\widehat{\mu}_i^{k,r} \;:=\; \frac{\frac{1}{N}\sum_{j=1}^{N}(Z_i^j)^r I_{\{\mathcal{M}(Z^j)\geq\widehat{\gamma}^k\}}\frac{f(Z^j)}{\widehat{g}^{k-1}(Z^j)}}{\widehat{\alpha}^k}. \tag{20}$$

Note that when $5'$ is used no extra care is required at the first iteration. In the numerical experiments of Section 4 we adopt both approaches, using step 5 for discrete distributions and step $5'$ for gamma distributions.

3.2. Discussion

We now discuss the extent to which Algorithm 1 differs from prior work on cross entropy (compiled in de Boer et al. 2005). The major differences can be classified into two categories: (i) update of the IS distributions, and (ii) update of the parameter ρ and of the sample size N.

One of the contributions of this paper is a generalization of the framework for the cross-entropy method, which allows for derivation of closed-form solutions to the CE-optimization problem for arbitrary distributions. This leads to a different update of the IS distributions, as reflected in step 5 of Algorithm 1. In the original CE method, the update of the IS distribution requires (i) working with parametric distributions, and (ii) solving a *stochastic optimization problem*, a task that can be very time-consuming. As discussed in de Boer et al.

(2005) (see also Homem-de-Mello and Rubinstein 2002), one case where this can be avoided is when the underlying distributions belong to the so-called natural exponential family, parameterized by the mean. In that case a closed-form solution to that stochastic optimization problem can be found — not surprisingly, the expression for the optimal parameter (the mean) coincides with (13) with $b(Z_i) = Z_i$. Such a class covers a good range of distributions but leaves out a number of cases, for example, discrete distributions or multi-parameter distributions such as normal or gamma (with both parameters allowed to vary). Algorithm 1 covers these cases by means of steps 5 or 5′, as illustrated numerically in Section 4.

In theory, the density function given by $\bar{g}_i$ in (10) is the best one can have under the cross-entropy philosophy, in the sense that any further constraints imposed to (4) — such as restricting g to be of parametric form — will yield sub-optimal solutions. In practice, of course, computing the whole CE-optimal density may be impractical; we have already discussed that, for continuous distributions, the approach of computing moments of the distribution (i.e. step 5′) provides an alternative to computing the CE-optimal density. A natural question that arises is how this moment-matching approximation performs for parametric distributions.

Consider the class of parametric distributions for which the parameters can be expressed as functions of the first, say, k moments. That is, suppose the original distribution f in (1) is of the form $f(z) = f_1(z_1, \theta_1) \times \cdots \times f_n(z_n, \theta_n)$, where the θ_i are parameters that can be written as functions of the first k moments. We represent this by writing $\theta_i = H_i(m_i^1, \ldots, m_i^k)$.

The original CE approach for such a problem will calculate the CE-optimal values of θ_i, call them $\tilde{\theta}_i$. Clearly, this is equivalent to finding the corresponding moments $\tilde{m}_i^1, \ldots, \tilde{m}_i^k$. Now suppose one applies the moment-matching approach described above. Then, one obtains the moments $\bar{m}_i^1, \ldots, \bar{m}_i^k$ of the CE-optimal density $\bar{g}_i$. How can we sample from a distribution with these moments? One natural alternative is to calculate $\bar{\theta}_i := H_i(\bar{m}_i^1, \ldots, \bar{m}_i^k)$ and then sample from $f_i(z_i, \bar{\theta}_i)$; in that case, it is clear that the quality of the $\bar{\theta}_i$ cannot be better than that of the $\tilde{\theta}_i$ obtained by optimizing directly the parameters, since in general $\bar{g}_i(\cdot)$ is not of the form $f_i(\cdot, \theta_i)$. Although the moment-matching approach will provide no better solutions than the original (parametric) CE method in this case, we remark that (i) the moments $\bar{m}_i^1, \ldots, \bar{m}_i^k$ can be used with distributions other than f_i, (ii) the moment-matching approach does not require solving a stochastic optimization problem, and (iii) in our experience, the resulting values of the parameters $\bar{\theta}_i$ and $\tilde{\theta}_i$ are in practice very similar. Indeed, it is reasonable to expect that the CE-optimal density $\bar{g}_i$ be close to the family of the original

distribution (say, a gamma distribution); hence, the moments of $\bar{g}_i$ should be close to the moments of the distribution obtained with the parametric approach. The results presented in Section 4 confirm that intuition.

The other major difference between the algorithm proposed in this paper and the original CE method refers to the update of the parameter ρ and of the sample size N in step 7. This step is crucial not only to establish that the algorithm terminates after finitely many iterations but also to provide a safeguard for practical convergence. Indeed, as discussed in Homem-de-Mello and Rubinstein (2002), one cannot expect Algorithm 1 to terminate if the parameter ρ is kept fixed throughout the algorithm. Roughly speaking, if ρ is fixed, the values of $\widehat{\gamma}^k$ may start converging to a value below the desired threshold a. Reducing ρ forces $\widehat{\gamma}^k$ to increase. If the random variable $\mathcal{M}(Z)$ satisfies certain conditions (e.g., if it has infinite tail), then one can guarantee that $\widehat{\gamma}^k$ can always be increased by a minimum amount, though in the process the sample size may have to be increased. Rubinstein (2002) also proposes an adaptive algorithm — where ρ is changed adaptively — but the motivation for the adaptive rules and hence the algorithm itself are different from the ones proposed here.

3.3. Convergence Issues

We formalize now the convergence notions discussed at the end of Section 3.2. We start with the following assumption:

Assumption A: The IS distributions selected by the algorithm belong to a class $\mathcal{G}$ such that $P_g(\mathcal{M}(Z) \geq a) > 0$ for all $g \in \mathcal{G}$.

Assumption A simply ensures that the probability being estimated — $P_f(\mathcal{M}(Z) \geq a)$ — does not vanish when the original pdf $f(\cdot)$ is replaced by a another distribution $g(\cdot)$. The assumption is trivially satisfied if the distribution of $\mathcal{M}(Z)$ has infinite tail when the distribution of Z belongs to some family (e.g., exponential, or gamma, etc.). For zero-tail distributions, the assumption holds as long as either a is less than the maximum value of the function $\mathcal{M}(Z)$, or if there is a positive probability that a is attained.

As before, let $\gamma(g, \rho)$ denote an arbitrary $(1-\rho)$-quantile of $\mathcal{M}(Z)$ under $g(\cdot)$. It is clear that, under assumption A, by decreasing ρ sufficiently we can force the quantiles γ to grow past a. In particular, we can force γ to increase at least by some pre-specified amount $\delta > 0$.

Thus, it is clear that $\gamma(\widehat{g}^{k-1}, \rho^{k-1}) \geq a$ for some k. However, the exact value of $\gamma(\widehat{g}^{k-1}, \rho^{k-1})$ is unknown; hence, we must ensure that such a property is kept when $\gamma(\widehat{g}^{k-1}, \rho^{k-1})$ is replaced by its estimate $\widehat{\gamma}_{N}(Z, \rho^{k-1})$.

Proposition 2 below does exactly that. In the proposition, the term "with probability one" refers to the probability space where Z lies, and when $Z^1, Z^2, \ldots$ are viewed as random variables on that space. Before stating the proposition, we show the lemma below, which is an interesting result in its own right since convergence results for quantiles found in the literature typically introduce an assumption to guarantee uniqueness of the quantiles (see, e.g., Serfling 1980). Lemma 1 shows that we still have convergence even when the quantile is not unique, though in that case one cannot guarantee convergence of sample quantiles to a single value.

Lemma 1 *Let $Y^1, Y^2, \ldots$ be i.i.d. random variables with common cdf $G(\cdot)$, and let Ξ denote the set of $(1-\rho)$-quantiles of G. Let $\widehat{\xi}_{N}$ denote a $(1-\rho)$-sample quantile of $Y^1, Y^2, \ldots, Y^N$. Then, the distance $d(\widehat{\xi}_{N}, \Xi)$ between $\widehat{\xi}_{N}$ and the set Ξ goes to zero (as N goes to infinity) with probability one. Moreover, given any $\varepsilon > 0$, we have $P(d(\widehat{\xi}_{N}, \Xi) > \varepsilon) \to 0$ exponentially fast with N.*

Proof. Notice initially that a $(1-\rho)$-quantile of a random variable Y can be expressed as an optimal solution of the problem $\min_{\xi} \mathbb{E}\phi(Y, \xi)$, where

$$\phi(Y, \xi) \;=\; \begin{cases} (1-\rho)(Y-\xi) & \text{if } \xi \leq Y \\ \rho(\xi - Y) & \text{if } \xi \geq Y. \end{cases}$$

To see this, notice that the subdifferential set $\partial_{\xi}\mathbb{E}\phi(Y, \xi)$ can be expressed as $\partial_{\xi}\mathbb{E}\phi(Y, \xi) = [\rho - P(Y \geq \xi),\ -(1-\rho) + P(Y \leq \xi)]$. It is easy to check that $\phi(Y, \xi)$ is convex in ξ for all Y. It follows that $\mathbb{E}\phi(Y, \xi)$ is convex in ξ and thus a necessary and sufficient optimality condition for the problem $\min_{\xi} \mathbb{E}\phi(Y, \xi)$ is $0 \in \partial\mathbb{E}\phi(Y, \xi)$ (see, e.g., Rockafellar 1970). This is true if and only if $\rho - P(Y \geq \xi) \leq 0$ and $-(1-\rho) + P(Y \leq \xi) \geq 0$, i.e., if and only if ξ is a $(1-\rho)$-quantile of Y. A similar argument shows that the sample $(1-\rho)$-quantile of a sample $Y_1, \ldots, Y_N$ (recall this is $\widehat{\xi}_{N}$) is the solution to the sample average approximation problem $\min_{\xi} N^{-1} \sum_{i=1}^{N} \phi(Y_i, \xi)$. Since the objective function $\mathbb{E}\phi(Y, \xi)$ is convex in ξ, it follows that the distance $d(\widehat{\xi}_{N}, \Xi)$ goes to zero as N goes to infinity w.p. 1 (Rubinstein and Shapiro 1993). The last statement follows from classical results on exponential rates of convergence of solutions of stochastic programs (Kaniovski et al. 1995). ■

Proposition 2 *Suppose assumption A holds, and let $x \in (0, a]$. Let $g \in \mathcal{G}$, and let $Z^1, Z^2, \ldots$ be i.i.d. with common density $g(\cdot)$. Then, there exists $\rho_x > 0$ and a random $N_x > 0$ such that, with probability one, $\widehat{\gamma}_{N}(Z, \rho) \geq x$ for all $\rho \in (0, \rho_x)$ and all $N \geq N_x$. Moreover, the probability that $\widehat{\gamma}_{N}(Z, \rho) \geq x$ for a given N goes to one exponentially fast with N.*

Proof. Let $\{Z^1, \ldots, Z^N\}$ be a set of i.i.d. samples from $g(\cdot)$. Consider initially the case where $P_g(\mathcal{M}(Z) > x) > 0$. As discussed earlier we have that $\gamma(g, \rho^*) > x$ for any $\rho^* \in (0, \rho_x^+)$, where $\rho_x^+ = P_g(\mathcal{M}(Z) > x) > 0$. It follows from Lemma 1 that the distance between the sample $(1-\rho^*)$-quantile $\widehat{\gamma}_{N}(Z, \rho^*)$ of $\mathcal{M}(Z^1), \ldots, \mathcal{M}(Z^N)$ and the set of $(1-\rho^*)$-quantiles of $\mathcal{M}(Z)$ goes to zero as N goes to infinity w.p. 1. Since $\gamma(g, \rho^*) > x$, it follows that $\widehat{\gamma}_{N}(Z, \rho^*) > x$ w.p. 1 for N large enough. Moreover, the probability that $\widehat{\gamma}_{N}(Z, \rho^*) > x$ for a given N goes to one exponentially fast.

Consider now the case where $P_g(\mathcal{M}(Z) > x) = 0$, i.e. x is the maximum value achieved by $\mathcal{M}(Z)$. By assumption A, this implies that $\rho_x^0 := P_g(\mathcal{M}(Z) = x) > 0$ and thus, for any $\rho^* \in (0, \rho_x^0)$ we must have $\gamma(g, \rho^*) = x$. It follows that $\gamma(g, \rho^*) = x$ is also the unique $(1-\rho^*)$-quantile of the random variable $W := x I_{\{\mathcal{M}(Z)=x\}}$. It is clear that $\widehat{\gamma}_{N}^{x} := x I_{\{\widehat{\gamma}_{N}(Z,\rho^*)=x\}}$ is a sample $(1-\rho^*)$-quantile of $W^1, \ldots, W^N$, where $W^j := x I_{\{\mathcal{M}(Z^j)=x\}}$. Since the distribution of W has finite support, it follows from the results in Shapiro and Homem-de-Mello (2000) that $\widehat{\gamma}_{N}^{x} = \gamma(g, \rho^*) = x$ w.p. 1 for N large enough, and, moreover, the probability that $\widehat{\gamma}_{N}^{x} = \gamma(g, \rho^*) = x$ for a given N goes to one exponentially fast. Since $\widehat{\gamma}_{N}^{x} = x$ if and only if $\widehat{\gamma}_{N}(Z, \rho^*) = x$, the proof is complete. ■

The above proposition shows not only that $\widehat{\gamma}_{N}(Z, \rho)$ reaches any threshold x for sufficiently small ρ and sufficiently large N (which ensures that the algorithm terminates), but also that one expects N not to be too large due to the exponential convergence, at least for moderate values of ρ (of course, when ρ is very small N needs to be large anyway). Notice that the update of the sample size in step 7(c) guarantees that the sample size N_x in Proposition 2 is achieved and hence either $\widehat{\gamma}^k$ increases by at least δ or it hits the value a. That is, at some iteration K we set $\widehat{\gamma}^K := a$. This ensures that Algorithm 1 finishes after a finite number of iterations. At that point we can then use the distribution $\widetilde{g}$ returned by the algorithm to calculate the estimate $\widehat{\alpha}_N(\widetilde{g})$ in (2), perhaps with a different sample size. Of course, $\widetilde{g}$ is only an estimate of the CE-optimal distribution $\bar{g}$; thus, the more one allows N to grow — which is controlled by the initial sample size as well as the update parameter ν — the more precise this estimate will be.

3.4. Practical Issues

Despite the above convergence results, one needs to be aware that, ultimately, the quality of the distributions generated by the algorithm will depend on the particular sample sizes used. A "poor" distribution will yield poor estimates of the underlying rare-event probabilities. Thus, it is important to ensure that "large enough" sample sizes are being used. Although in general such a calculation is problem-dependent — and as such must be determined by experimentation — in some cases it is possible to derive some guidelines.

For example, consider the case of estimating an arbitrary function of each random variable. Using (13), one can easily construct estimates using sample average counterparts. For example, the moments of each random variable can be estimated by (20). The ratio form of that expression suggests use of a procedure to calculate confidence intervals originally developed for estimation of ratios (e.g., regenerative simulation). Following that approach, in order to obtain a $(1-\beta)$%-confidence interval for $\theta_i := \mathbb{E}_{\bar{g}_i}[b(X_i)]$ in (13) we can draw a set $\{Z^1, \ldots, Z^N\}$ of i.i.d. samples from some $g(\cdot)$ (in case of Algorithm 1, $g = \widehat{g}^{k-1}$), calculate

$$\begin{aligned} \widehat{\alpha} &:= \frac{1}{N}\sum_{j=1}^{N} \mathcal{N}(Z^j)\frac{f(Z^j)}{g(Z^j)} \\ \widehat{\theta}_i &:= \left(\frac{1}{N}\sum_{j=1}^{N} b(Z_i^j)\mathcal{N}(Z^j)\frac{f(Z^j)}{g(Z^j)}\right)\frac{1}{\widehat{\alpha}} \end{aligned}$$

and then the interval is given by

$$\widehat{\theta}_i \pm \frac{z_{1-\beta/2}\sqrt{\widehat{\sigma}^2/N}}{\widehat{\alpha}}.$$

In the above formulas, $z_{1-\beta/2}$ is the standard normal $(1-\beta)$-quantile and $\widehat{\sigma}^2 := \widehat{\sigma}_{11} - 2\widehat{\theta}_i\widehat{\sigma}_{12} + \widehat{\theta}_i^2\widehat{\sigma}_{22}$, where the $\widehat{\sigma}_{ij}$ are the elements of the sample covariance matrix Σ that estimates the covariance between $b(Z_i)\mathcal{N}(Z)\frac{f(Z)}{g(Z)}$ and $\mathcal{N}(Z)\frac{f(Z)}{g(Z)}$. Having confidence intervals for θ_i as a function of the sample size allows us to control the error of the estimates by computing the appropriate sample sizes. Such a procedure is standard in simulation; see Law and Kelton (2000) for details.

Another issue related to practical implementation of Algorithm 1 concerns the values for the constants ν, δ, and ρ^0. We suggest $\nu \leq 2$. For δ, one approach is to take $\delta = 0$ until the sequence $\{\widehat{\gamma}^k\}$ gets "stalled," at which point a positive δ is used again. This approach yields the slowest progression of $\widehat{\gamma}^k$, but in turn the final estimate $\widehat{g}^k$ is more "reliable,"

in the sense that it is calculated from more precise estimates — recall from Section 3 that we need the event $\{\mathcal{M}(Z) \geq \gamma^k\}$ not to be rare under $\widehat{g}^{k-1}$. This also explains why it is desirable that ρ^k does not become too small; otherwise large sample sizes will be required to get reliable estimates $\widehat{g}^k$. Notice however that, even if the CE-optimal $\bar{g}$ could be obtained, some problems might still require a very large sample size in (2). Given the limitations of one's computational budget, Algorithm 1 can be used to detect such a situation — the algorithm can be halted once ρ^k in step 7 gets too small (or, equivalently, when N gets too large).

4. Numerical Results

We present now numerical results obtained for a manufacturing problem. The model is described in the manufacturing setting only to simplify the discussion — as we shall see below, the problem may be cast in terms of longest paths in a certain directed graph. General longest-path models are widely applicable in many areas, for example PERT networks (Adlakha and Kulkarni 1989). Since the proposed algorithm does not exploit the structure of the graph when computing the IS distributions, we believe the behavior of the algorithm for a more general longest-path problem would be similar to the one observed here.

In all examples below, we used an implementation of Algorithm 1 described in Section 3.3. Recall that the algorithm requires the definition of three constants, ρ_0, ν, and δ. We used $\rho_0 = 0.1$ and $\nu = 2$. For δ, we adopted the conservative approach $\delta = 0$. In these examples such a δ sufficed, i.e., the sequence $\{\widehat{\gamma}^k\}$ never got stalled. We also implemented a control of the sample sizes as a function of the error of the estimates $\widehat{\theta}_i$, as discussed in Section 3.4. More specifically, let Δ_i be the ratio between the half-width of the confidence interval for θ_i and the estimate $\widehat{\theta}_i$. Our goal was to keep the maximum (with respect to i) value of Δ_i constant at all iterations, although we allowed N to grow by a factor of at most ν per iteration.

Consider a single stage in a production system in which there are S 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 s is a random variable whose distribution is known, and that each station processes its incoming 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 s ($s = 1, ..., S$), let Y_{sj} denote the service time of processing job j on station s, and let C_{sj} denote the time job j finishes its service at station s. Let $Y := (Y_{11}, \ldots, Y_{SJ})$ denote the vector of service times, which is assumed to be random with a known distribution. Note that C_{Sj} can be viewed as a total *completion time* of job j and that each C_{sj} is a function of Y, and hence is random. The above model was studied in Homem-de-Mello et al. (1999) in the context of optimizing the performance system with respect to the release times of the jobs, so no estimation of probabilities was involved; 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 a; that is, with $\mathcal{M}(Y) = C_{SJ}(Y)$, we want to estimate $\alpha = P(\mathcal{M}(Y) \geq a)$. Calculation of $\mathcal{M}(Y)$ for a particular realization of Y can be done via the recursive formula

$$C_{sj} = \max(C_{s-1,j},\ C_{s,j-1}) + Y_{sj}, \quad j = 1, \ldots, J, \quad s = 1, \ldots, S, \tag{21}$$

with $C_{s0} = C_{0j} = 0,\ s = 1, \ldots, S,\ j = 1, \ldots, J$. Notice that the above problem is *static* (which is the focus of the present paper) since the number of jobs under consideration is finite.

The structure of the problem allows for derivation of *lower and upper bounds* for the probability of interest. For that, we shall use the fact that C_{SJ} is the length of the longest path in a certain directed graph, which is a "grid" of $S + 1 \times J$ nodes except that the nodes on the last row are not connected horizontally (see Homem-de-Mello et al. 1999 for a detailed description). Let T denote the total number of feasible paths and L_p the length of path p, $p = 1, \ldots, T$. Thus, $C_{SJ} = \max_{p=1,\ldots,T} L_p$.

It is easy to see that each feasible path has exactly $S + J - 1$ arcs — more specifically, S vertical arcs and $J - 1$ horizontal ones. Since each arc length corresponds to a service time, it follows that each L_p is the sum of $S + J - 1$ random variables. The proposition below gives a useful property if all jobs have the same distribution on a given machine. Notice that it deals with the concept of *stochastic ordering* (denoted by $\geq_{st}$); we refer to Asmussen (2003) for definitions and properties. The proof of the proposition is provided in the Online Supplement to this paper on the journal's website.

Proposition 3 *Suppose that, for each $s = 1, \ldots, S$, the random variables $\{Y_{sj}\}_{j=1,\ldots,J}$ are identically distributed, and that all random variables in the problem are independent. Suppose also that one of the random variables $\{Y_{s1}\}_{s=1,\ldots,S}$ dominates the others in the stochastic*

sense, i.e., there exists $s_{max} \in \{1, \ldots, S\}$ *such that* $Y_{s_{max}1} \geq_{st} Y_{s1}$, $s = 1, \ldots, S$. *Then, there exists a path* p_{max} *such that* $L_p \leq_{st} L_{p_{max}}$ *for all* $p = 1, \ldots, T$. *Moreover,*

$$P(L_{p_{max}} \geq x) \;\leq\; P(C_{SJ} \geq x) \;\leq\; \binom{S+J-2}{J-1} P(L_{p_{max}} \geq x). \tag{22}$$

In particular, if the random variables $\{Y_{s1}\}_{s=1,\ldots,S}$ *are such that* $Y_{s1} \geq_{st} Y_{\ell 1}$ *if and only if* $\mathbb{E}[Y_{s1} \geq Y_{\ell 1}]$ *then* p_{max} *denotes the index of the path with the largest* expected *value.*

4.1. Gamma Distributions

We consider the case where all service times have a gamma distribution. In what follows, we denote by gamma(η, β) the gamma distribution with mean $\eta\beta$ and variance $\eta\beta^2$. Clearly, the parameters η and β can be recovered from the first two moments of the distribution, since

$$\eta\beta^2 \;=\; \mathrm{Var}[Y] \;\iff\; \mathbb{E}[Y]\beta \;=\; \mathbb{E}[Y^2] - (\mathbb{E}[Y])^2 \;\iff\; \beta \;=\; \frac{\mathbb{E}[Y^2]}{\mathbb{E}[Y]} - \mathbb{E}[Y] \tag{23}$$

$$\eta\beta \;=\; \mathbb{E}[Y] \;\iff\; \eta \;=\; \frac{\mathbb{E}[Y]}{\beta} \tag{24}$$

For simplicity, we assume that the service times of all jobs are independent, and that the service times of all jobs at a given machine have the same distribution. Even with this simplifying assumption, exact computation of α is impossible, except for special cases. Thus, in our view the example provides a meaningful and novel application of the proposed algorithm.

We adopted the following methodology. First, for fixed J and S, we generated a problem randomly. This was accomplished by generating parameters $\eta_1, \ldots, \eta_S$ uniformly between 1 and 5 and $\beta_1, \ldots, \beta_S$ uniformly between 1 and 10 (one pair (η_s, β_s) for each machine). We then estimated $P(C_{SJ} \geq a)$ for three values of a, based on the value of the total mean service time $\Gamma = J \sum_{s=1}^{S} \eta_s \beta_s$. We took $a = 0.8\Gamma$, $a = \Gamma$, and $a = 2\Gamma$. The rationale for these choices was that the expected completion time would be Γ if a job started its process at machine 1 only after the previous job finished its process at the last machine. Thus, Γ is a gross overestimate of the actual expected completion time, hence $P(C_{SJ} \geq \Gamma)$ should be small.

To estimate $P(C_{SJ} \geq a)$, we used Algorithm 1 to estimate the first two moments of the CE-optimal distribution, and then recovered the optimal parameters η^* and β^* using (23)-(24). The output of the algorithm — two $S \times J$-dimensional vectors — determined the parameters of the gamma importance-sampling distribution used to estimate the probability.

For the sake of comparison, we also estimated the same probability using the *hazard-rate twisting* (HRT) method described in Juneja and Shahabuddin (2002) and Huang and Shahabuddin (2004). The HRT method consists of twisting the original distributions by an exponential quantity that depends on the so-called asymptotic hazard function of the distribution. More specifically, let $\Lambda(x) = -\log(1 - F(x))$ denote the hazard function of a distribution F, and let $\widetilde{\Lambda}(x)$ denote a function that is asymptotically similar to $\Lambda(x)$. The HRT method computes the twisted distribution

$$dF^*(x) \;=\; e^{\theta_a \widetilde{\Lambda}(x) - \Psi(\theta_a)} dF(x), \tag{25}$$

where θ_a is a carefully selected value that depends on the threshold value a, and $\Psi(\theta_a)$ is the normalization constant $\log \int e^{\theta_a \widetilde{\Lambda}(x)} dF(x)$.

In the present case of gamma distributions and for the particular function $\mathcal{M}(Y) = C_{SJ}(Y)$, the calculations are greatly simplified. Indeed, the asymptotic hazard function of a gamma(η, β) distribution is $\widetilde{\Lambda}(x) = (1/\beta)x$. Moreover, the normalization function $\Psi(\theta)$ is given by $-\eta \log(1 - \theta)$. To simplify the notation, let $Y_1, \ldots, Y_n$ denote the random variables in the problem (so $n = S \times J$), with $Y_i \sim$ gamma(η_i, β_i). Following Huang and Shahabuddin (2004), define the function $q(a) := ca$ for some $c > 0$. This function satisfies Condition 4.4 in Huang and Shahabuddin (2004), thus we have that $\lim_{a\to\infty} \log P(C_{SJ} > a)/q(a) = -I_{\text{opt}}$, where

$$I_{\text{opt}} \;=\; \inf \left\{ \sum_{i=1}^{n} y_i/\beta_i \;:\; C_{SJ}(y) > 1/c \right\}. \tag{26}$$

Since C_{SJ} is defined by max and + operations, it follows that the optimization problem in (26) can be easily solved. Let $\beta_{\max} := \max_{i=1,\ldots,n} \beta_i$ and let $i_{\max}$ be an index such that $\beta_{\max} = \beta_{i_{\max}}$. Then, the solution to the optimization problem in (26) is simply $y_i^* = 1/c$ if $i = i_{\max}$ and $y_i^* = 0$ otherwise. It follows that $I_{\text{opt}} = 1/(c\beta_{\max})$ and hence

$$P(C_{SJ} > a) \;=\; e^{-a/\beta_{\max}(1+o(1))}. \tag{27}$$

Huang and Shahabuddin (2004) suggest then taking $\theta_a = 1 - b/q(a) = 1 - b/(ca)$ for some $b > 0$. By substituting this value into (25) we see that the twisted distribution F_i^* for variable Y_i is given by

$$F_i^* \;=\; \text{gamma}\left(\eta_i, \frac{\beta_i ca}{b}\right), \tag{28}$$

whose meaning should be clear despite the abuse of notation. Unfortunately, despite the asymptotic optimality of the distribution in (28), the performance of this distribution for

finite values of a seems to depend very much on the choice of the constants c and b — indeed, in our examples this distribution often performed very poorly and yielded estimates equal to zero in several cases. A possible explanation is that the HRT procedure hinges on $\widetilde{\Lambda}(x)$ being a good approximation for $\Lambda(x)$. While this is true asymptotically, the approximation may be poor even for large values of x, particularly if the number of random variables is relatively large.

Nevertheless, the idea of hazard-rate twisting is very appealing, so in order to use that method we chose values for c and b in an empirical way. Our rationale was the following: from (28) we have that the mean of Y_i under the twisted distribution is $\eta_i \beta_i ca/b$. We would like the paths defining the completion times C_{SJ} to have total mean equal to a. Moreover, we know from (27) that the maximum value among $\beta_1, \ldots, \beta_n$ is what defines the asymptotic probability. Thus, we chose b so that the completion time calculated with weights $\eta_i \beta_{\max} ca/b$ on the arcs is equal to a. This corresponds to taking

$$b \;=\; \sum_{i \in p^0} \eta_i \beta_{\max} c, \tag{29}$$

where p^0 is the path corresponding to that completion time. Note that the above value is similar to the one proposed by Juneja et al. (2004), which can be shown to be equal to $\sum_{i=1}^{n} \eta_i \beta_{\max} c$ (though the latter did not perform well in our experiments). In either case, when b is substituted into (28) the constant c disappears from the expression.

To provide a fair comparison, we provided the same *computational budget* for both methods. That is, we used a larger sample size for HRT, since Algorithm 1 requires extra computational time to calculate the optimal parameters. We increased the sample size sequentially until the total CPU time used by the HRT method was the same as the time used for Algorithm 1. Based on these samples, we computed the estimates for mean and variance. As an extra verification, the above procedure was replicated ten times, and we built 95% confidence intervals using the averages of individual estimates of mean and variance.

We also compared our algorithm with the parametric CE method, in which the parameters are optimized directly. As discussed in Section 3.2, optimizing the parameters usually requires solving a difficult stochastic optimization problem; in the particular case of gamma distributions, however, the calculations are simplified. Indeed, in that case the CE problem to be solved is

$$\max_{\eta_i, \beta_i \geq 0} \quad \int \log[f_i(z_i, \eta_i, \beta_i)]\, \varphi_i(z_i)\, dz_i \tag{30}$$

where $f_i(z_i, \eta_i, \beta_i) = \Gamma(\eta_i)^{-1}\beta_i^{-\eta_i} z_i^{\eta_i - 1} e^{-z_i/\beta_i}$ is the density of the gamma distribution and $\varphi_i(z_i)$ is defined in (6). After some algebra, one can easily re-write the objective function of the above problem as

$$\psi_i(\eta_i, \beta_i) \;:=\; \alpha \left[\log \frac{\beta_i^{-\eta_i}}{\Gamma(\eta_i)} + (\eta_i - 1) \frac{\mathbb{E}_f \left[\log(Z_i) I_{\{\mathcal{M}(Z) \geq a\}} \right]}{\alpha} - \frac{1}{\beta} \frac{\mathbb{E}_f \left[Z_i I_{\{\mathcal{M}(Z) \geq a\}} \right]}{\alpha} \right]. \quad (31)$$

Note that the right-most term is exactly the same as expression (14) for the first moment of the CE-optimal density $\bar{g}_i$. Thus, we can estimate this value using the same multi-stage procedure given by Algorithm 1 — and a slight modification of the algorithm also allows for estimation of $\mathbb{E}_f \left[\log(Z_i) I_{\{\mathcal{M}(Z) \geq a\}} \right] / \alpha$. Note also that we can divide the objective function by α since we are interested only in the optimal solution of (30). It follows that, once the expectations in (31) are estimated, (30) becomes a simple deterministic two-dimensional problem, which can be solved using standard optimization methods. We used Matlab's `fminsearch` function, which in turn implements the Nelder-Mead algorithm.

Confirming the intuitive argument laid out in Section 3.2, the values of η_i and β_i obtained with the parametric procedure described above were very similar to the values obtained with the moment-matching approach (see Table 4 for one example). We emphasize, however, that the latter method does not require the extra optimization step — which, even though it takes negligible time in this particular case, may be difficult for other distributions. The estimated probabilities with both methods were in most cases statistically equal; for that reason, we do not display the results obtained with the parametric approach.

Another possible way to bypass the optimization procedure is to allow only one of the parameters (say, β_i) to vary; in that case, the procedure becomes closer to the versions of the CE method proposed in the literature for distributions in the natural exponential family, where the optimal mean can be estimated directly. Clearly, such a procedure can only provide sub-optimal solutions with respect to the approach where both η_i and β_i are optimized; for example, for the system whose results are displayed in Table 4, the variance of the one-parameter estimate was about three times as large as the variance of the two-parameter estimate. Thus, we do not report results for the one-parameter approach.

Table 1 displays the results for $J = 10$ jobs and $S = 5$ machines, which corresponds to 50 random variables. The values of the parameters η_i and β_i for this data set are respectively 2, 4, 5, 5, 3 and 10, 8, 1, 10, 9 for each machine. Although these results correspond to a particular instance of data, similar results were observed for other problems we generated

(for the same J and S and the same rule for generation of η and β). Therefore, we report only one representative of the group. In the table, $\widehat{\alpha}$ is the estimate for $P(C_{SJ} \geq a)$, with the number in parentheses denoting the half-width of a 95% confidence interval, in the same order of magnitude. N is the sample size used to compute the estimate once the importance-sampling distribution is determined. Notice that, since the sample size used with the HRT method was variable, the N column under "HRT" displays a rounded average. The column labelled "N_{CE}" reports the (rounded) average sample size used in Algorithm 1 to calculate the CE-optimal parameters (recall that we used an adaptive scheme for automatic update of sample sizes, as described earlier). The initial sample size used in the procedure was always set to 5000.

To compute the bounds given by (22), we need to estimate $P(L_{p_{max}} \geq a)$. For the underlying problem, one can easily check from the construction in the proof of Proposition 3 that $L_{p_{max}} \stackrel{d}{=} \sum_{i=1}^{10} \text{gamma}(5,10) + \text{gamma}(2,10) + \text{gamma}(4,8) + \text{gamma}(5,1) + \text{gamma}(3,9)$ (note that gamma$(5,10)$ stochastically dominates the other distributions). We thus obtain the bounds

$$P\left(\text{gamma}(50,10) \geq a\right) \;\leq\; P\left(L_{p_{max}} \geq a\right) \;\leq\; P\left(\text{gamma}(70,10) \geq a\right). \tag{32}$$

Table 1 lists the exact lower and upper bounds given by (22) and (32). Notice that $T = \binom{13}{9} = 715$ in that case. Exact solutions for such a problem are not available, so the purpose of the bounds is just to provide a rough check on the order of magnitude of the obtained results. Note that HRT underestimates the probability when $a = 1340$ and $a = 2680$.

Table 1: Estimated Probabilities and Exact Bounds for the Case $J = 10$, $S = 5$, $\eta_s \sim U(1,5)$, $\beta_s \sim U(1,10)$

	HRT		CE				
a	$\widehat{\alpha}$	N	$\widehat{\alpha}$	N	N_{CE}	lower bound	upper bound
1072	$4.8\ (6.8) \times 10^{-8}$	127K	$9.0\ (0.1) \times 10^{-8}$	100K	13K	2.5×10^{-10}	3.8×10^{-2}
1340	$2.9\ (2.7) \times 10^{-18}$	158K	$4.7\ (0.1) \times 10^{-14}$	100K	26K	2.8×10^{-17}	3.2×10^{-7}
2680	$2.3\ (3.9) \times 10^{-98}$	322K	$8.6\ (0.1) \times 10^{-56}$	100K	80K	7.8×10^{-61}	7.9×10^{-45}

We also studied the case where all service times have the same gamma distribution with parameters $\eta = 1$ and $\beta = 25$. In that case, the bounds in (22) can be computed more precisely since $L_{p_{max}}$ is the sum of $S + J - 1$ gamma$(1,25)$ independent random variables and thus has a gamma$(14,25)$ distribution. Table 2 displays the estimation results, together with the bounds given by (22). Note that HRT underestimates the probability when $a = 2500$.

Table 2: Estimated Probabilities and Exact Bounds for the Case $J = 10$, $S = 5$, $\eta_s = 1$, $\beta_s = 25$

	HRT		CE				
a	$\widehat{\alpha}$	N	$\widehat{\alpha}$	N	N_{CE}	lower bound	upper bound
1000	$1.2\ (0.4) \times 10^{-5}$	117K	$5.9\ (0.3) \times 10^{-5}$	100K	10K	6.7×10^{-7}	4.8×10^{-4}
1250	$2.8\ (4.7) \times 10^{-8}$	160K	$3.7\ (0.9) \times 10^{-8}$	100K	32K	5.1×10^{-10}	3.6×10^{-7}
2500	$3.7\ (7.2) \times 10^{-32}$	348K	$9.5\ (5.9) \times 10^{-28}$	100K	80K	6.9×10^{-28}	4.9×10^{-25}

Finally, we studied the case where $S = 1$. In this case, the completion time is simply a sum of J i.i.d. gamma(η, β) and therefore has a gamma$(J\eta, \beta)$ distribution, so the probabilities can be computed analytically. In this case we took $a = 2\Gamma$, $a = 3\Gamma$, and $a = 4\Gamma$, where $\Gamma = J\eta\beta$. Table 3 displays the results for $J = 10$, $\eta = 5$, and $\beta = 5$. The column labeled "Exact" contains the true values. We can see that the estimates obtained with both methods are very close to the real values, and indeed the confidence intervals cover the exact values. This suggests that the heuristics we used to determine the parameters of the HRT method is efficacious, at least when the number of variables is small.

Table 3: Estimated and Exact Probabilities for the Case $J = 10$, $S = 1$, $\eta = 5$, $\beta = 5$

	HRT		CE			
a	$\widehat{\alpha}$	N	$\widehat{\alpha}$	N	N_{CE}	Exact
500	$1.176\ (0.008) \times 10^{-8}$	123K	$1.178\ (0.007) \times 10^{-8}$	100K	5K	1.179×10^{-8}
750	$7.401\ (0.055) \times 10^{-22}$	161K	$7.390\ (0.072) \times 10^{-22}$	100K	26K	7.412×10^{-22}
1000	$1.704\ (0.013) \times 10^{-37}$	225K	$1.696\ (0.020) \times 10^{-37}$	100K	44K	1.693×10^{-37}

To illustrate the behavior of the algorithm, we considered another problem with $J = 5$ jobs, $S = 2$ machines, and all service times having the same gamma$(5, 5)$ distribution. The value of a chosen was $a = 350$, for which the algorithm yielded the estimate probability $1.006(0.049) \times 10^{-7}$ with sample size 100,000 (the lower and upper bounds for this case are respectively 2.433×10^{-8} and 1.216×10^{-7}). Table 4 displays, for each iteration k, the value of $\widehat{\gamma}^k$ (computed in step 2 of Algorithm 1), the corresponding sample size used, and the new parameters η and β of each service time Y_{sj}, obtained from the moments calculated in step 5′ of the algorithm. Notice that $\widehat{\gamma}^k$ reaches $a = 350$ after 4 iterations. The last line displays the values obtained by solving the parametric problem (30) with the objective function re-written as in (31), which yielded the estimate $1.014(0.060) \times 10^{-7}$. For the same problem (and same

computational budget), the HRT method yielded the estimate $1.335(0.522) \times 10^{-7}$, with the IS distribution being a gamma$(5, 11.67)$.

Table 4: Progression of the Algorithm for the Case $J = 5$, $S = 2$, $\eta = 5$, $\beta = 5$

k	$\widehat{\gamma}^k$	N	$\widehat{\eta}^k$					$\widehat{\beta}^k$				
0		5K	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00
			5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00
1	200.8	5K	6.02	5.32	4.67	4.14	4.25	5.61	6.21	6.49	6.97	6.91
			4.36	4.17	4.28	4.68	6.00	6.47	7.04	7.09	6.88	5.39
2	255.6	5K	5.92	4.34	4.03	4.10	3.59	7.36	8.93	8.74	8.06	8.43
			3.59	3.08	4.48	4.34	6.25	8.64	10.80	8.08	8.88	6.76
3	310.3	10K	6.19	4.44	3.48	2.63	3.61	8.72	10.26	11.94	14.30	8.07
			2.84	2.66	2.95	4.00	6.78	10.82	13.38	13.63	11.53	8.11
4	350.0	20K	5.56	4.28	3.36	2.43	3.38	10.39	12.90	13.39	16.64	8.84
			2.61	2.53	2.94	4.13	5.95	12.20	15.53	15.69	12.87	9.80
parametric			5.15	3.69	3.23	2.61	3.68	11.23	15.00	13.93	15.50	8.12
CE			3.19	2.82	3.00	3.62	5.71	10.00	13.93	15.36	14.66	10.21

4.2. Discrete Distributions

We now consider the case where all service times have discrete distributions with finite support. As before, we assume that the service times of all jobs are independent, and that the service times of all jobs at a given machine have the same distribution.

For fixed J, S, and m we generated, for each of the S machines, m values for service times between 10 and 40 and m corresponding probabilities at random. 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. However, such a procedure does not determine the probability of the maximum value, unless there is a single path corresponding to it. We then estimated $P(C_{SJ} \geq a)$ for two values of a, based on the value of the maximum completion time Ψ. We took $a = 0.9\Psi$ and $a = \Psi$ (obviously, $P(C_{SJ} > \Psi) = 0$).

To estimate $P(C_{SJ} \geq a)$, we again used Algorithm 1. In this case we can determine the whole IS distribution, which reduces to the probabilities of each value of each service time — an $S \times J \times m$-dimensional vector. In order to check the obtained probabilities, we also estimated the same values using standard Monte Carlo, providing the same computational budget for both methods. The above procedure was replicated 50 times, and the average and

95% confidence intervals were built from those 50 independent estimates, both for Monte Carlo and CE.

We first consider a problem with $J = 10$ jobs, $S = 5$ machines, and $m = 4$ possible outcomes for each random variable. This corresponds to 50 random variables and a 200-dimensional parameter vector. The values $y_{s\ell}$ taken on by each service time and the respective probabilities $p_{s\ell}$ are listed in Table 5. In this particular case the exact probability for $a = \Psi = 541$ can be computed, since there is a single path corresponding to the maximum completion time. That value is $(0.330)(0.392)^{10}(0.466)(0.220)(0.197) = 5.710 \times 10^{-7}$. The estimated probabilities are displayed in Table 6, using notation similar to Tables 1-3. We can see that the estimate obtained with the CE method is fairly close to the real value.

Table 5: Values Taken on by the Service Times and Corresponding Probabilities, for the Data Set with Discrete Distributions

s	y_{s1}	p_{s1}	y_{s2}	p_{s2}	y_{s3}	p_{s3}	y_{s4}	p_{s4}
1	12	0.309	16	0.091	28	0.270	39	0.330
2	11	0.035	25	0.418	32	0.155	40	0.392
3	16	0.137	17	0.353	28	0.044	38	0.466
4	17	0.635	20	0.037	21	0.108	29	0.220
5	18	0.679	20	0.072	23	0.052	35	0.197

Table 6: Estimated Probabilities for Discrete-Distribution Case, $J = 10$, $S = 5$, $m = 4$, Random Data

	MC		CE			
a	$\widehat{\alpha}$	N	$\widehat{\alpha}$	N	N_{CE}	Exact
486	$2.30\ (0.18) \times 10^{-2}$	635	$2.22\ (0.21) \times 10^{-2}$	100	100	
541	0.000 (0.00)	10200	$4.95\ (3.26) \times 10^{-7}$	700	700	5.71×10^{-7}

To illustrate the behavior of the algorithm for the discrete-distribution case, we consider a smaller problem with $J = 4$ jobs and $S = 3$ machines; the distribution of the service times is the same as in the first three rows of Table 5. The maximum value achieved by C_{SJ} in this case is $a = 237$, for which the algorithm yielded the estimate probability 0.0035 (± 0.0002) with 20 replications of sample size 200 each (the exact value can be calculated as 0.0036). Table A–1 in the Online Supplement displays, for each iteration k, the value of $\widehat{\gamma}^k$ (computed in step 3 of Algorithm 1) and the updated probability of each value taken on by Y_{sj}, as calculated in step 5 of the algorithm (denoted by $\widehat{p}^k_{sj,m}$). Notice that $\widehat{\gamma}^k$ reaches

a after three iterations. Notice also the presence of a "degenerate" effect; the (s, j) with 1.0 in the respective row correspond to the edges of the longest path in the related graph. Incidentally, this example illustrates the application of the CE method to combinatorial optimization problems (in this case, longest path). We refer to de Boer et al. (2005) and Rubinstein (1999, 2002) for details.

5. Concluding Remarks

We have studied some aspects of the cross-entropy method, which is an algorithm for estimation of rare-event probabilities that has been proposed in the literature and that has been gaining some popularity. More specifically, we have proposed a general form of the method — applicable to any distribution — that encompasses and extends existing work. We have also proposed an implementable version of the algorithm and illustrated its behavior through numerical examples. The obtained results are encouraging and suggest that the proposed algorithm is fairly robust, requiring little tuning of its parameters.

Some issues for further research remain. For example, it would be important to find conditions under which the solutions of the cross-entropy and variance-minimization problems coincide, at least asymptotically. Also, the derivation of performance bounds for the estimates obtained with the proposed method (derived with finite sample size) would be desirable.

Acknowledgments

We are grateful to Reuven Rubinstein for introducing the CE method to us and for his comments on early versions of this paper. We also thank the Area Editor and two referees for comments that considerably improved the presentation of our results.

References

Adlakha, V. G., V. G. Kulkarni. 1989. A classified bibliography of research on stochastic PERT networks: 1966-1987. *INFOR* **27** 272–296.

Asmussen, S. 2003. *Applied Probability and Queues*. 2nd ed. Springer-Verlag, Berlin, Germany.

Bonnans, J. F., A. Shapiro. 2000. *Perturbation Analysis of Optimization Problems*. Springer Series in Operations Research, Springer-Verlag, New York.

Bucklew, J. A. 1990. *Large Deviation Techniques in Decision, Simulation, and Estimation*. Wiley, New York.

de Boer, P.T. 2000. Analysis and efficient simulation of queueing models of telecommunications systems. Ph.D. thesis, Department of Computer Science, Univ. of Twente, The Netherlands.

de Boer, P. T., D. P. Kroese, S. Mannor, R. Y. Rubinstein. 2005. A tutorial on the cross-entropy method. *Ann. of Oper. Res.* **134** 19–67.

Dembo, A., O. Zeitouni. 1998. *Large Deviations Techniques and Applications*. 2nd ed. Springer-Verlag, New York.

Fishman, G. 1997. *Monte Carlo: Concepts, Algorithms and Applications*. Springer-Verlag, New York.

Glasserman, P. 2004. *Monte Carlo Methods in Financial Engineering*. Applications of Mathematics, vol. 53. Springer-Verlag, New York.

Glasserman, P., P. Heidelberger, P. Shahabuddin, T. Zajic. 1999. Multilevel splitting for estimating rare event probabilities. *Oper. Res.* **47** 585–600.

Glynn, P. W., D. L. Iglehart. 1989. Importance sampling for stochastic simulations. *Management Sci.* **35** 1367–1392.

Heidelberger, P. 1995. Fast simulation of rare events in queueing and reliability models. *ACM Trans. on Model. and Comput. Simulation* **5** 43–85.

Homem-de-Mello, T., R. Y. Rubinstein. 2002. Estimation of rare event probabilities using cross-entropy. E. Yücesan, C.-H. Chen, J. L. Snowdon, J. M. Charnes, eds. *Proc. of the 2002 Winter Simulation Conf.* 310–319.

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

Huang, Z., P. Shahabuddin. 2004. A unified approach for finite dimensional, rare-event Monte Carlo simulation. R. G. Ingalis, M. D. Rossetti, J. S. Smith, B. A. Peters, eds. *Proc. of the 2004 Winter Simulation Conf.* 1616–1624.

Juneja, S., R. Karandikar, P. Shahabuddin. 2004. Tail asymptotes and fast simulation of delay probabilities in stochastic PERT networks. Manuscript, Tata Institute of Fundamental Research, Mumbai, India.

Juneja, S., P. Shahabuddin. 2002. Simulating heavy tailed processes using delayed hazard rate twisting. *ACM Trans. on Model. and Comput. Simulation* **12** 94–118.

Kahn, H., A. W. Marshall. 1953. Methods of reducing the sample size in Monte Carlo computations. *J. of the Oper. Res. Soc.* **1** 263–278.

Kaniovski, Y. M., A. J. King, R. J.-B. Wets. 1995. Probabilistic bounds (via large deviations) for the solutions of stochastic programming problems. *Ann. Oper. Res.* **56** 189–208.

Kapur, J. N., H. K. Kesavan. 1992. *Entropy Optimization Principles with Applications*. Academic Press, New York.

Kroese, D., R. Y. Rubinstein. 2004. The transform likelihood ratio method for rare event simulation with heavy tails. *Queueing Systems* **46** 317–351.

Kullback, S., R. A. Leibler. 1951. On information and sufficiency. *Ann. Math. Statist.* **22** 79–86.

Law, A. M., W. D. Kelton. 2000. *Simulation Modeling and Analysis*. 3rd ed. McGraw-Hill, New York.

Oh, M.-S., J. O. Berger. 1992. Adaptive importance sampling in Monte Carlo integration. *J. Statist. Comput. Simulation* **41** 143–168.

Oh, M.-S., J. O. Berger. 1993. Integration of multimodal functions by Monte Carlo importance sampling. *J. Amer. Statist. Assoc.* **88** 450–456.

Rockafellar, R. T. 1970. *Convex Analysis*. Princeton Univ. Press, Princeton, NJ.

Rubinstein, R. Y. 1999. The cross-entropy method for combinatorial and continuous optimization. *Methodology and Comput. in Appl. Probab.* **2** 127–190.

Rubinstein, R. Y. 2002. Cross-entropy and rare events for maximal cut and partition problems. *ACM Trans. on Model. and Comput. Simulation* **12** 27–53.

Rubinstein, R. Y., A. Shapiro. 1993. *Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization by the Score Function Method*. Wiley, Chichester, U.K.

Serfling, R. 1980. *Approximation Theorems in Mathematical Statistics*. Wiley, New York.

Shahabuddin, P. 1995. Rare event simulation of stochastic systems. C. Alexopoulos, K. Kang, W. R. Lilegdon, D. Goldsman, eds. *Proc. of the 1995 Winter Simulation Conf.* 178–185.

Shapiro, A., T. Homem-de-Mello. 2000. On the rate of convergence of Monte Carlo approximations of stochastic programs. *SIAM J. on Optim.* **11** 70–86.

Vázquez-Abad, F., D. Dufresne. 1998. Accelerated simulation for pricing Asian options. D. J. Medeiros, E. F. Watson, J. S. Carson, M. S. Manivannan, eds. *Proc. of the 1998 Winter Simulation Conf.* 1493–1500.

Villén-Altamirano, M., J. Villén-Altamirano. 1999. About the efficiency of RESTART. *Proc. of the RESIM Workshop*. Univ. of Twente, The Netherlands 99–128.

Zhang, P. 1996. Nonparametric importance sampling. *J. Amer. Statist. Assoc.* **91** 1245–1253.