

Rare Event Probability Estimation for Static Models via Cross-Entropy and Importance Sampling

Tito Homem-de-Mello

Department of Industrial, Welding and Systems Engineering

The Ohio State University

1971 Neil Ave., Columbus, Ohio 43210-1271, USA

e-mail: homem-de-mello.1@osu.edu

Reuven Y. Rubinstein

Faculty of Industrial Engineering and Management

Technion—Israel Institute of Technology

Haifa 32000, Israel

e-mail: ierrr01@ie.technion.ac.il

July 5, 2002

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. In theory, probabilities of events can be estimated using a simple procedure — draw independent and identically distributed samples from the underlying distributions, and compute the fraction of occurrences of the event under study. This is called a *crude Monte Carlo* technique.

The above approach, although simple, is only practical when the probabilities being estimated are not very small. For *rare* events, it requires a prohibitively large numbers of trials in most interesting cases. Thus, new techniques are required. Given its importance, this problem has attracted considerable attention from the simulation research community. Among the methods developed are the *splitting/RESTART* (see, for instance, [7, 8, 11, 14, 29]) and *importance sampling* techniques (see, e.g., [10]). Here we focus on importance sampling (IS) techniques.

The main idea of IS, when applied to rare events, is to make their occurrence more frequent, or in other words, to “speed up” the simulation. Technically, IS aims to select a probability distribution (change of measure) that minimizes the variance of the IS estimator. 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 finding the right change of measure is described by results based on large deviations theory. This type of analysis yields efficient algorithms, but is usually feasible only for relatively simple models; see [2, 13, 19, 27] for surveys. Thus, IS has only been successfully applied to systems with relative low complexity.

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, [25, 26]. 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 [22], 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. It was soon realized [23, 24] that a simple modification of [22] (involving minimization of the so-called *Kullback-Leibler cross-entropy* with respect to the tilted parameter, instead of minimization of the variance) could be

used not only for estimating probabilities of rare events but also as a powerful heuristic method for difficult combinatorial optimization problems. De Boer [4] in his PhD thesis presents several efficient heuristics based on state-dependent exponential changes of measure to overcome the difficulties where the state-independent cross-entropy method fails.

In the rare-events context, the Kullback-Leibler cross-entropy is used to define a “distance” between the IS distribution and the (unknown) optimal zero-variance measure. We can then find the parameter that minimizes this quantity. 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, the obtained parameter is asymptotically close to the parameter that minimizes variance.

In this paper we expand the work initiated in [22] and concentrate on the application of the cross-entropy method (henceforth called CE method) to estimate rare event probabilities in *static* models. We feel that this class of problems is wide enough to justify our focus on it. Indeed, any *transient* problem can be viewed that way. For example, many problems in manufacturing fall into that category (see section 6 for an example); also, many problems in finance — an area that has received great attention from the simulation community, see e.g. [9] — also have a static nature. In addition, as mentioned above static rare event problems have an interesting connection with combinatorial optimization. An application of the CE method to dynamic systems such as queueing networks is given in [5].

The main purpose of this paper is to establish a foundation for the use of CE method to estimate rare event probabilities in static models. After reviewing some concepts and describing the basic technique, we study its relation to the aforementioned variance minimization problem in the parametric setting. Our discussion suggests that, as the underlying events become rarer, the optimal parameter given by the CE method tends to coincide with the parameter that minimizes variance. We then discuss several examples, where the computations can be performed analytically, that corroborate that notion.

We also provide an adaptive scheme to estimate the optimal CE-parameters. The procedure, which involves a sequence of stochastic optimization subproblems, can be applied quite generally. It works particularly well if the underlying distributions have *finite support* or if they belong to the so-called *natural exponential family* (NEF), since in those cases there are analytical solutions to those subproblems — see section 3.

The convergence of the proposed method is discussed in section 5. We establish that the adaptive procedure we introduce converge to an estimate of the optimal CE-parameter after *finitely many* iterations and with a *finite* sample size. The estimate obtained can then be refined by increasing the sample size if necessary.

Finally, we present in section 6 some numerical results to illustrate the method. The model we study is 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. The results suggest that the CE method works very well, in that it provides accurate estimates for probabilities as low as 10^{-27} in reasonable time. These results are checked through the derivation of lower and upper bounds, or even exact values in some cases. We also show that crude Monte Carlo fails to estimate such probabilities, even when the same computational budget is allocated.

2 Background on Importance Sampling and Cross-Entropy

2.1 Importance Sampling

To set up the stage, we review some background material from [23]. Let ℓ be the expected performance of a stochastic system given in the form

$$\ell := \mathbb{E}_f[\varphi[\mathcal{M}(\mathbf{Y})]] = \int \varphi[\mathcal{M}(\mathbf{y})] f(\mathbf{y}) d\mathbf{y}, \quad (2.1)$$

where $\mathcal{M}(\mathbf{Y})$ is the *sample performance*, the subscript f in $\mathbb{E}_f[\varphi[\mathcal{M}(\mathbf{Y})]]$ means that the expectation of the random vector \mathbf{Y} is taken with respect to the probability density function (pdf) f which has a light tail density (e.g. gamma, Poisson, truncated, etc. [25]), and $\varphi[\cdot]$ is a real valued function. Examples of $\varphi[\mathcal{M}(\mathbf{y})]$ are $\varphi[\mathcal{M}(\mathbf{y}, x)] = I_{\{\mathcal{M}(\mathbf{y}) > x\}}$ and $\varphi[\mathcal{M}(\mathbf{y}, \gamma)] = \exp\left(-\frac{\mathcal{M}(\mathbf{y})}{\gamma}\right)$, where x and γ are given constants. As an example of $\mathcal{M}(\cdot)$, consider the shortest path in a stochastic network, which can be defined as $\mathcal{M}(\mathbf{Y}) = \min_{j=1, \dots, p} \sum_{i \in \mathcal{B}_j} Y_i$. Here, \mathcal{B}_j is the j -th complete path from a source to a sink; p is the number of complete paths; and $Y_i, i = 1, \dots, m$, are the durations of the i -th component. For the longest path, of course, we replace min with max in the definition of \mathcal{M} .

Let $G(\mathbf{y})$ be a probability measure (distribution) such that $dG(\mathbf{y}) = g(\mathbf{y})d\mathbf{y}$, where $g(\mathbf{y})$ is a pdf. Assume that $g(\mathbf{y})$ dominates $\varphi[\mathcal{M}(\mathbf{y})]f(\mathbf{y})$ in the absolutely continuous sense, that is, $\text{supp}\{\varphi[\mathcal{M}(\mathbf{y})]f(\mathbf{y})\} \subset \text{supp}\{g(\mathbf{y})\}$, 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 ℓ as

$$\ell = \int \varphi[\mathcal{M}(\mathbf{z})] \frac{f(\mathbf{z})}{g(\mathbf{z})} g(\mathbf{z}) d\mathbf{z} = \mathbb{E}_g \left[\varphi[\mathcal{M}(\mathbf{Z})] \frac{f(\mathbf{Z})}{g(\mathbf{Z})} \right],$$

where the subscript g means that the expectation is taken with respect to g , which is called the *importance sampling* (IS) pdf.

An unbiased estimator of ℓ is

$$\hat{\ell}_N = \frac{1}{N} \sum_{i=1}^N \varphi[\mathcal{M}(\mathbf{Z}_i)] W(\mathbf{Z}_i), \quad (2.2)$$

where $W(\mathbf{z}) = f(\mathbf{z})/g(\mathbf{z})$ is called the *likelihood ratio* (LR), and $\mathbf{Z}_1, \dots, \mathbf{Z}_N$ are independent and identically distributed (i.i.d.) samples from $g(\mathbf{z})$. In the particular case where there is no change of measure ($g = f$), we have $W = 1$, and the LR estimator $\widehat{\ell}_N$ reduces to the following so-called *crude Monte Carlo* (CMC) estimator:

$$\tilde{\ell}_N = \frac{1}{N} \sum_{i=1}^N \varphi[\mathcal{M}(\mathbf{Y}_i)] ,$$

where $\mathbf{Y}_1, \dots, \mathbf{Y}_N$ are i.i.d. samples from the pdf $f(\mathbf{y})$.

The choice of the dominating pdf $g(\mathbf{y})$ is crucial for the variance of the LR estimator (2.2). We consider next the problem of minimizing the variance of $\widehat{\ell}_N$ with respect to the pdf g , that is,

$$\min_g \text{Var}_g \left[\varphi[\mathcal{M}(\mathbf{Z})] \frac{f(\mathbf{Z})}{g(\mathbf{Z})} \right]. \quad (2.3)$$

It is well known (e.g. [25]) that the solution of problem (2.3) is

$$g^*(\mathbf{z}) = \frac{|\varphi[\mathcal{M}(\mathbf{z})]| f(\mathbf{z})}{\int |\varphi[\mathcal{M}(\mathbf{z})]| f(\mathbf{z}) d\mathbf{z}}. \quad (2.4)$$

Note that if $\varphi[\mathcal{M}(\mathbf{Z})] \geq 0$, then

$$g^*(\mathbf{z}) = \frac{\varphi[\mathcal{M}(\mathbf{z})] f(\mathbf{z})}{\ell} \quad (2.5)$$

and $\text{Var}[\widehat{\ell}_N] = 0$. The density $g^*(\mathbf{z})$ as per (2.4) and (2.5) is called the *optimal importance sampling density*.

In general, however, implementation of the optimal importance sampling pdf $g^*(\mathbf{z})$ as per (2.4) and (2.5) is problematic. The main difficulty lies in the fact that in order to derive $g^*(\mathbf{z})$ one needs to know ℓ . But ℓ is precisely the quantity we want to estimate from the simulation! In most simulation studies the situation is even worse, since the analytical expression for the sample performance is unknown in advance. To overcome this difficulty, one can make a pilot run with the underlying model, obtain a sample $\mathcal{M}(\mathbf{Y}_1), \dots, \mathcal{M}(\mathbf{Y}_N)$, and then use it to estimate $g^*(\mathbf{z})$. It is important to note that sampling from such an artificially constructed pdf $g(\mathbf{z})$ might be a very complicated and time-consuming task, especially when $g(\mathbf{z})$ is a high-dimensional pdf.

An alternative approach to the above problem can be derived when the underlying pdf's belong to some parametric family $\mathcal{F} = \{f(\mathbf{y}, \mathbf{v}), \mathbf{v} \in V\}$. *Throughout this paper, we will assume that this is the case.* Let $f(\mathbf{y}, \mathbf{u})$ denote the pdf of the random vector \mathbf{Y} in (2.1). 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 $f(\mathbf{y}) = f(\mathbf{y}, \mathbf{u})$ by a single parameter (vector) \mathbf{v} . For more details see [25]. The likelihood ratio W in (2.2) with $g(\mathbf{y}) = f(\mathbf{y}, \mathbf{v})$ reduces to $W(\mathbf{Z}, \mathbf{u}, \mathbf{v}) = f(\mathbf{Z}, \mathbf{u})/f(\mathbf{Z}, \mathbf{v})$, where \mathbf{v} ($\mathbf{v} \neq \mathbf{u}$) is called the *reference* parameter vector. In this case the program (2.3) reduces to

$$\min_{\mathbf{v} \in V} \text{Var}_{\mathbf{v}} [\varphi[\mathcal{M}(\mathbf{Z})] W(\mathbf{Z}, \mathbf{u}, \mathbf{v})]. \quad (2.6)$$

It is readily seen that the optimal solution of (2.6) coincide with that of

$$\min_{\mathbf{v} \in V} \left\{ \mathcal{V}(\mathbf{u}, \mathbf{v}) := \mathbb{E}_{\mathbf{v}} \left[(\varphi[\mathcal{M}(\mathbf{Z})])^2 (W(\mathbf{Z}, \mathbf{u}, \mathbf{v}))^2 \right] \right\}. \quad (2.7)$$

The above problem can still be difficult to solve, since the pdf with respect to which the expectation is computed depends on the decision variable \mathbf{v} . To overcome this obstacle, we rewrite (2.7) as

$$\min_{\mathbf{v} \in V} \left\{ \mathcal{V}(\mathbf{u}, \mathbf{v}) := \mathbb{E}_{\mathbf{v}_1} \left[(\varphi[\mathcal{M}(\mathbf{X})])^2 W(\mathbf{X}, \mathbf{u}, \mathbf{v}) W(\mathbf{X}, \mathbf{u}, \mathbf{v}_1) \right] \right\}. \quad (2.8)$$

Note that (2.8) is obtained from (2.7) by multiplying and dividing the integrand by $f(\mathbf{x}, \mathbf{v}_1)$. Note also that in (2.7) and (2.8) the expectation is taken with respect to the pdf's $f(\cdot, \mathbf{v})$ and $f(\cdot, \mathbf{v}_1)$, respectively. Moreover, $W(\mathbf{X}, \mathbf{u}, \mathbf{v}_1) = f(\mathbf{X}, \mathbf{u})/f(\mathbf{X}, \mathbf{v}_1)$, and $\mathbf{X} \sim f(\mathbf{x}, \mathbf{v}_1)$. Note finally that for the particular case $\mathbf{v}_1 = \mathbf{u}$ we obtain from (2.8) that

$$\min_{\mathbf{v} \in V} \left\{ \mathcal{V}(\mathbf{u}, \mathbf{v}) := \mathbb{E}_{\mathbf{u}} \left[(\varphi[\mathcal{M}(\mathbf{Y})])^2 W(\mathbf{Y}, \mathbf{u}, \mathbf{v}) \right] \right\}, \quad (2.9)$$

where as before $\mathbf{Y} \sim f(\mathbf{y}, \mathbf{u})$. We shall call (2.9) the *variance-minimization* problem.

Clearly, the optimal solution of the programs (2.8)- (2.9), say \mathbf{v}^* , is typically not available analytically, since the expected value $\ell(\mathbf{v}) = \mathbb{E}_{\mathbf{v}}[\varphi[\mathcal{M}(\mathbf{Y})]]$ cannot be evaluated exactly. To overcome this difficulty, we can replace the expected value by a Monte Carlo estimate, that is, solve

$$\min_{\mathbf{v} \in V} \left\{ \hat{\mathcal{V}}_N(\mathbf{u}, \mathbf{v}) := N^{-1} \sum_{i=1}^N (\varphi[\mathcal{M}(\mathbf{Y}_i)])^2 W(\mathbf{Y}_i, \mathbf{u}, \mathbf{v}) \right\}, \quad (2.10)$$

(where $\mathbf{Y}_1, \dots, \mathbf{Y}_N$ are i.i.d. samples generated from the original pdf $f(\mathbf{y}, \mathbf{u})$), and then take the optimal solution of (2.10) as an estimator of \mathbf{v}^* . This type of procedure has been well studied in the literature, appearing under various names such as *stochastic counterpart method*, *sample-path optimization* or *sample average approximation*. Under proper assumptions, it is possible to show that the optimal solutions of (2.10) converge to \mathbf{v}^* as N goes to infinity. See, for instance, [25, 26] for more details.

We now borrow from [25] a simple recursive algorithm (see Algorithm 8.4.1 in [25]), which is based on the stochastic counterpart of (2.8), i.e., on

$$\min_{\mathbf{v} \in V} \left\{ \hat{\mathcal{V}}_N(\mathbf{u}, \mathbf{v}) := N^{-1} \sum_{i=1}^N \varphi^2[\mathcal{M}(\mathbf{X}_i)] W(\mathbf{X}_i, \mathbf{u}, \mathbf{v}) W(\mathbf{X}_i, \mathbf{u}, \mathbf{v}_1) \right\}. \quad (2.11)$$

Algorithm 2.1 :

1. Set $\mathbf{v}_1 = \mathbf{u}$. Generate a sample $\mathbf{X}_1, \dots, \mathbf{X}_N$ from the pdf $f(\mathbf{y}, \mathbf{v}_1)$ and solve the stochastic program (2.11). Denote the solution by $\hat{\mathbf{v}}^*$. Take $\hat{\mathbf{v}}^*$ as an estimate of the optimal reference parameter \mathbf{v}^* .

2. Estimate the performance measure ℓ using the LR estimate (2.2) with $g(\mathbf{z})$ replaced by $f(\mathbf{z}, \hat{\mathbf{v}}^*)$.

To get a more accurate estimate of \mathbf{v}^* one can set $\mathbf{v}_1 = \hat{\mathbf{v}}_1^*$ and repeat step 1 for several additional iterations.

2.2 The Cross-Entropy Method

An alternative way for estimating the optimal reference parameter vector \mathbf{v}^* is based on the Kullback–Leibler *cross-entropy* [18], which defines a “distance” between the two probability distributions (densities) $f(\mathbf{y})$ and $g(\mathbf{y})$ and can be written as

$$\mathcal{D}(f, g) = \int f(\mathbf{y}) \ln \frac{f(\mathbf{y})}{g(\mathbf{y})} d\mathbf{y}. \quad (2.12)$$

Notice that \mathcal{D} is not a distance in the formal sense, since in general $\mathcal{D}(f, g) \neq \mathcal{D}(g, f)$. Also, although the above expression involves pdf’s — and we keep this terminology throughout the paper — we emphasize that relation (2.12) holds for discrete distributions as well, with probability mass functions (pmf) in place of pds’s and summations in place of integrals.

Clearly, when $g = f$ we have $\mathcal{D}(f, g) = 0$. Note that both cross-entropy and the variance minimization problem (2.3) are particular cases of the Ali–Silvey “distance” [1] between two probability densities, which is defined as

$$d(f, g) = \aleph \left[\int f(\mathbf{y}) \psi \left(\frac{f(\mathbf{y})}{g(\mathbf{y})} \right) d\mathbf{y} \right],$$

where $\aleph(\cdot)$ is a continuous convex function on $(0, +\infty)$ and $\psi(\cdot)$ is an increasing, real-valued function of a real variable. Indeed, in the case of the cross-entropy and variance minimization with $g(\mathbf{y}) = c^{-1}|\varphi[\mathcal{M}(\mathbf{y})]|f(\mathbf{y})$, ($c = \int \varphi[\mathcal{M}(\mathbf{y})]f(\mathbf{y})d\mathbf{y}$) we have $\psi(y) = \ln(y)$, $\aleph(y) = y$ and $\psi(y) = y^2$, $\aleph(y) = y$, respectively.

To proceed, let $g(\mathbf{y}) = f(\mathbf{y}, \mathbf{v})$ and assume that $\varphi[\mathcal{M}(\mathbf{y})] \geq 0$. Noting that $\phi(\mathbf{y}, \mathbf{v}) := c^{-1}\varphi[\mathcal{M}(\mathbf{y})]f(\mathbf{y}, \mathbf{v})$ (with $c = \int \varphi[\mathcal{M}(\mathbf{y})]f(\mathbf{y}, \mathbf{v})d\mathbf{y}$) is a pdf, we can define a cross-entropy between $\phi(\mathbf{y}, \mathbf{u})$ and $f(\mathbf{y}, \mathbf{v})$, in analogy to (2.12), as

$$\begin{aligned} \mathcal{D}(\mathbf{u}, \mathbf{v}) &:= \int c^{-1}\varphi[\mathcal{M}(\mathbf{y})]f(\mathbf{y}, \mathbf{u}) \ln \frac{c^{-1}\varphi[\mathcal{M}(\mathbf{y})]f(\mathbf{y}, \mathbf{u})}{f(\mathbf{y}, \mathbf{v})} d\mathbf{y} \\ &= \mathbb{E}_{\mathbf{u}} \left[c^{-1}\varphi[\mathcal{M}(\mathbf{Y})] \ln \frac{c^{-1}\varphi[\mathcal{M}(\mathbf{Y})]f(\mathbf{Y}, \mathbf{u})}{f(\mathbf{Y}, \mathbf{v})} \right], \end{aligned}$$

and find the reference parameter vector \mathbf{v}^* that solves the following problem

$$\min_{\mathbf{v} \in V} \left\{ \mathcal{D}(\mathbf{u}, \mathbf{v}) := \mathbb{E}_{\mathbf{u}} \left[c^{-1}\varphi[\mathcal{M}(\mathbf{Y})] \ln \frac{c^{-1}\varphi[\mathcal{M}(\mathbf{Y})]f(\mathbf{Y}, \mathbf{u})}{f(\mathbf{Y}, \mathbf{v})} \right] \right\}. \quad (2.13)$$

It is obvious that the optimal solutions of (2.13) and of $\max_{\mathbf{v} \in V} \{D(\mathbf{u}, \mathbf{v}) := \mathbb{E}_{\mathbf{u}} [\varphi[\mathcal{M}(\mathbf{Y})] \ln f(\mathbf{Y}, \mathbf{v})]\}$ are equivalent. Using LR we can also write the latter problem in analogy to (2.11) as

$$\max_{\mathbf{v} \in V} \{D(\mathbf{u}, \mathbf{v}) := \mathbb{E}_{\mathbf{v}_1} [\varphi[\mathcal{M}(\mathbf{X})] W(\mathbf{X}, \mathbf{u}, \mathbf{v}_1) \ln f(\mathbf{X}, \mathbf{v})]\}. \quad (2.14)$$

Given a sample $\mathbf{X}_1, \dots, \mathbf{X}_N$ from $f(\mathbf{x}, \mathbf{v}_1)$, we can estimate the optimal solution \mathbf{v}^* of (2.14) by the optimal solution of the program

$$\max_{\mathbf{v} \in V} \left\{ \widehat{D}_N(\mathbf{u}, \mathbf{v}) := \frac{1}{N} \sum_{i=1}^N \varphi[\mathcal{M}(\mathbf{X}_i)] W(\mathbf{X}_i, \mathbf{u}, \mathbf{v}_1) \ln f(\mathbf{X}_i, \mathbf{v}) \right\}. \quad (2.15)$$

Note that the programs (2.14) and (2.15), which are based on CE, can be considered as alternatives to the variance minimization problems (2.9) and (2.10) respectively.

2.3 Estimation of Rare Event Probabilities

When not stated otherwise, we shall consider in the sequel the case where the performance measure ℓ has the form

$$\ell(x) := P(\mathcal{M}(\mathbf{Y}) \geq x) = \mathbb{E}_{\mathbf{u}} [I_{\{\mathcal{M}(\mathbf{Y}) \geq x\}}], \quad (2.16)$$

(that is, $\ell(x)$ represents the probability of the rare-event $\{\mathcal{M}(\mathbf{Y}) \geq x\}$). For convenience of reference, we rewrite our basic formulae (2.2), (2.8), (2.11), (2.14) and (2.15):

$$\widehat{\ell}_N(x, \mathbf{v}_1) := \frac{1}{N} \sum_{i=1}^N I_{\{\mathcal{M}(\mathbf{X}_i) \geq x\}} W(\mathbf{Z}_i, \mathbf{u}, \mathbf{v}_1), \quad (2.17)$$

$$\min_{\mathbf{v} \in V} \left\{ \mathcal{V}(\mathbf{u}, \mathbf{v}) := \mathbb{E}_{\mathbf{v}_1} [I_{\{\mathcal{M}(\mathbf{X}) \geq x\}} W(\mathbf{X}, \mathbf{u}, \mathbf{v}) W(\mathbf{X}, \mathbf{u}, \mathbf{v}_1)] \right\}, \quad (2.18)$$

$$\min_{\mathbf{v} \in V} \left\{ \widehat{\mathcal{V}}_N(\mathbf{u}, \mathbf{v}) := \frac{1}{N} \sum_{i=1}^N I_{\{\mathcal{M}(\mathbf{X}_i) \geq x\}} W(\mathbf{X}_i, \mathbf{u}, \mathbf{v}) W(\mathbf{X}_i, \mathbf{u}, \mathbf{v}_1) \right\}, \quad (2.19)$$

$$\max_{\mathbf{v} \in V} \left\{ D(\mathbf{u}, \mathbf{v}) := \mathbb{E}_{\mathbf{v}_1} [I_{\{\mathcal{M}(\mathbf{X}) \geq x\}} W(\mathbf{X}, \mathbf{u}, \mathbf{v}_1) \ln f(\mathbf{X}, \mathbf{v})] \right\}, \quad (2.20)$$

$$\max_{\mathbf{v} \in V} \left\{ \widehat{D}_N(\mathbf{u}, \mathbf{v}) := \frac{1}{N} \sum_{i=1}^N I_{\{\mathcal{M}(\mathbf{X}_i) \geq x\}} W(\mathbf{X}_i, \mathbf{u}, \mathbf{v}_1) \ln f(\mathbf{X}_i, \mathbf{v}) \right\}, \quad (2.21)$$

respectively, where as before $W(\mathbf{X}, \mathbf{u}, \mathbf{v}) = f(\mathbf{X}, \mathbf{u})/f(\mathbf{X}, \mathbf{v})$ and $\mathbf{X}_1, \dots, \mathbf{X}_N$ are i.i.d. samples from the pdf $f(\mathbf{z}, \mathbf{v}_1)$.

The theoretical framework in which one typically examines rare-event probability estimation is based on complexity theory, as introduced in [20]. There, IS estimators are classified either

as *polynomial-time* or as *exponential-time*. It is shown in [25] that for an IS estimator, $\widehat{\ell}_N(x)$ of $\ell(x)$, to be polynomial-time, it suffices that its *squared coefficient of variation* (SCV), $\kappa^2(x)$, also called relative error, be bounded in x by some polynomial function, $p(x)$. For such polynomial-time estimators, the required sample size to achieve a fixed relative error does not grow too fast as the event becomes rarer. In order to obtain such (polynomially bounded) $\kappa^2(x) \equiv \kappa^2(\mathbf{v}, x)$, one minimizes $\kappa^2(\mathbf{v}, x)$ with respect to \mathbf{v} , where

$$\kappa^2(\mathbf{v}, x) = \frac{N\text{Var}[\widehat{\ell}_N(x, \mathbf{v})]}{\ell^2(x)}. \quad (2.22)$$

2.4 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 (2.4). The VM method ensures, by construction, the best approximation within the family $\{f(\mathbf{z}, \mathbf{v}), \mathbf{v} \in V\}$ — 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. In this section we provide a somewhat informal comparison for the case of estimation of rare-event probabilities.

Consider the VM and CE problems in the form (2.18) and (2.20), respectively, with $\mathbf{v}_1 = \mathbf{u}$. It is clear that we can replace the objective function $D(\mathbf{u}, \mathbf{v})$ in (2.20) by

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

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

$$\begin{aligned} \mathcal{V}(\mathbf{u}, \mathbf{v}) &= \mathbb{E}\mathbf{u} \left[I_{\{\mathcal{M}(\mathbf{Y}) \geq x\}} W(\mathbf{Y}, \mathbf{u}, \mathbf{v}) \right] \\ &= \mathbb{E}\mathbf{u} [W(\mathbf{Y}, \mathbf{u}, \mathbf{v}) | \mathcal{M}(\mathbf{Y}) \geq x] P_{\mathbf{u}}(\mathcal{M}(\mathbf{Y}) \geq x) \end{aligned} \quad (2.23)$$

and

$$\begin{aligned} D_1(\mathbf{u}, \mathbf{v}) &= -\mathbb{E}\mathbf{u} \left[I_{\{\mathcal{M}(\mathbf{Y}) \geq x\}} \ln W(\mathbf{Y}, \mathbf{u}, \mathbf{v}) \right] \\ &= -\mathbb{E}\mathbf{u} [\ln W(\mathbf{Y}, \mathbf{u}, \mathbf{v}) | \mathcal{M}(\mathbf{Y}) \geq x] P_{\mathbf{u}}(\mathcal{M}(\mathbf{Y}) \geq x). \end{aligned} \quad (2.24)$$

Notice the similarity between (2.23) and (2.24). Let now \mathbf{v}^* be an optimal solution to the VM problem. Thus, we must have $\mathcal{V}(\mathbf{u}, \mathbf{v}^*) - \mathcal{V}(\mathbf{u}, \mathbf{v}) \leq 0$ for all $\mathbf{v} \in V$, i.e.

$$\mathbb{E}\mathbf{u} \left[W(\mathbf{Y}, \mathbf{u}, \mathbf{v}) \frac{f(\mathbf{Y}, \mathbf{v}) - f(\mathbf{Y}, \mathbf{v}^*)}{f(\mathbf{Y}, \mathbf{v}^*)} \Big| \mathcal{M}(\mathbf{Y}) \geq x \right] \leq 0 \quad \text{for all } \mathbf{v} \in V. \quad (2.25)$$

On the other hand, if \mathbf{v}^* is an optimal solution to the CE problem then we must have

$$\mathbb{E}_{\mathbf{u}} \left[\ln \frac{f(\mathbf{Y}, \mathbf{v})}{f(\mathbf{Y}, \mathbf{v}^*)} \mid \mathcal{M}(\mathbf{Y}) \geq x \right] \leq 0 \quad \text{for all } \mathbf{v} \in V. \quad (2.26)$$

The solution sets defined by (2.25) and (2.26) are in general different. Suppose however that there exists \mathbf{v}^* such that $f(\mathbf{y}, \mathbf{v}^*) \geq f(\mathbf{y}, \mathbf{v})$ for all \mathbf{y} such that $\mathcal{M}(\mathbf{y}) \geq x$ and all $\mathbf{v} \in V$. It is clear that such \mathbf{v}^* satisfies both (2.25) and (2.26), i.e., such \mathbf{v}^* is both VM- and CE-optimal. This suggests that, as x goes to infinity — i.e. as $P_{\mathbf{u}}(\mathcal{M}(\mathbf{Y}) \geq x)$ goes to zero — the VM and CE problems tend to have the same solutions. The examples in section 4 corroborate that intuitive notion.

3 Basic Algorithm

We discuss now ways to solve the CE problem (2.20). When there exists an optimal solution of (2.20) in the *interior* of the set V , we can find stationary points of $D(\mathbf{u}, \cdot)$ by equating the gradient $\nabla_{\mathbf{v}} D(\mathbf{u}, \mathbf{v})$ to zero. Moreover, in case the expectation and differentiation operators can be interchanged, the stationary points of (2.20) and (2.21) can be obtained from the solution of the following nonlinear equations:

$$\nabla D(\mathbf{u}, \mathbf{v}) = \mathbb{E}_{\mathbf{v}_1} \left[I_{\{\mathcal{M}(\mathbf{X}) \geq x\}} W(\mathbf{X}, \mathbf{u}, \mathbf{v}_1) \nabla \ln f(\mathbf{X}, \mathbf{v}) \right] = \mathbf{0}$$

and

$$\nabla \hat{D}_N(\mathbf{u}, \mathbf{v}) = \frac{1}{N} \sum_{i=1}^N I_{\{\mathcal{M}(\mathbf{X}_i) \geq x\}} W(\mathbf{X}_i, \mathbf{u}, \mathbf{v}_1) \nabla \ln f(\mathbf{X}_i, \mathbf{v}) = \mathbf{0}.$$

Sufficient conditions under which the above interchange of operators is valid are: either (i) the function $\ln f(\mathbf{X}, \mathbf{v})$ is convex (or concave) in \mathbf{v} w.p.1, or (ii) the function $\ln f(\mathbf{X}, \mathbf{v})$ is Lipschitz with an integrable constant, that is, there exists a random variable M such that $\mathbb{E}M < \infty$ and, for any $\mathbf{v}, \mathbf{w} \in V$, $|\ln f(\mathbf{X}, \mathbf{v}) - \ln f(\mathbf{X}, \mathbf{w})| \leq M \|\mathbf{v} - \mathbf{w}\|$ w.p.1. We refer to [26] for details. The interchange is also valid, of course, if the distribution of \mathbf{X} has finite support, since in that case the expectation is a sum of finitely many terms.

3.1 Natural Exponential Family

A particular case of interest occurs when the components of the random vector $\mathbf{Y} = (Y_1, \dots, Y_n)$ are independent and each has a distribution in the *natural exponential family* (NEF) (see, e.g., [12]). A random variable X is said to have a NEF distribution if

$$f(y, w) = \exp(yw - k(w))h(y), \quad w \in W \subset \mathbb{R}, \quad (3.1)$$

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 can be written as particular cases of the expression above; see [25] for details. For example, for a Poisson pmf we have $f(y, \lambda) = \lambda^y e^{-\lambda} / y!$, which can be obtained from (3.1) by setting $w = \ln \lambda$, $k(w) = e^w$, $h(y) = 1/y!$ for $y \geq 0$ (y integer) and $h(y) = 0$ for $y < 0$.

It is possible to show that, if X has density $f(y, w)$ as in (3.1), then we have

$$\mu = \mathbb{E}X = k'(w) \quad (3.2)$$

$$\text{Var}[X] = k''(w). \quad (3.3)$$

We then re-parameterize (3.1) as $\tilde{f}(y, \mu) = \exp(yw(\mu) - k(w(\mu)))h(y)$, 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 $\text{Var}[X] > 0$. Notice that, when $\text{Var}[X] = 0$, i.e. when $X = \mu$ with probability one (w.p.1), we have from (3.2) that $k(w) = \mu w$.

The proposition below shows that the CE problem (2.20) can be solved exactly for NEF distributions.

Proposition 3.1 *Suppose that Y_1, \dots, Y_n are independent random variables such that Y_j has NEF distribution (3.1), that is $\tilde{f}_j(y_j, u_j) = \exp(y_j w_j(u_j) - k_j(w_j(u_j)))h_j(y_j)$. In addition, suppose that $P_{\mathbf{u}}(\mathcal{M}(\mathbf{Y}) \geq x) > 0$. Then, the maximizer \mathbf{v}^* of (2.20) over $V \equiv \mathbb{R}^n$ is given by*

$$v_j^* = \frac{\mathbb{E}_{\mathbf{u}} [Y_j I_{\{\mathcal{M}(\mathbf{Y}) \geq x\}}]}{\mathbb{E}_{\mathbf{u}} [I_{\{\mathcal{M}(\mathbf{Y}) \geq x\}}]}} = \frac{\mathbb{E}_{\mathbf{v}_1} [X_j W(\mathbf{X}, \mathbf{u}, \mathbf{v}_1) I_{\{\mathcal{M}(\mathbf{X}) \geq x\}}]}{\mathbb{E}_{\mathbf{v}_1} [I_{\{\mathcal{M}(\mathbf{X}) \geq x\}}] W(\mathbf{X}, \mathbf{u}, \mathbf{v}_1)}}. \quad (3.4)$$

Proof. It is easy to see that, under the independence assumption of the proposition, problem (2.20) (with $\mathbf{v}_1 = \mathbf{u}$) becomes separable and thus it reduces to the unidimensional problem

$$\begin{aligned} & \max_{v_j \in \mathbb{R}} \mathbb{E}_{\mathbf{u}} \left[I_{\{\mathcal{M}(\mathbf{Y}) \geq x\}} [\log h_j(Y_j) + Y_j w_j(v_j) - k_j(w_j(v_j))] \right] \\ &= \max_{v_j \in \mathbb{R}} \left\{ D_j(v_j) := w_j(v_j) \mathbb{E}_{\mathbf{u}} [Y_j I_{\{\mathcal{M}(\mathbf{Y}) \geq x\}}] - k_j(w_j(v_j)) \mathbb{E}_{\mathbf{u}} [I_{\{\mathcal{M}(\mathbf{Y}) \geq x\}}] \right\}. \end{aligned} \quad (3.5)$$

If $\text{Var}[Y_j] = 0$, i.e. if $Y_j = u_j$ w.p.1, then clearly $D_j(v_j) \equiv 0$ and so any v_j solves (3.5). Suppose now that $\text{Var}[Y_j] > 0$. Notice that, using (3.2) and (3.3), we have

$$\begin{aligned} D_j'(v_j) &= w_j'(v_j) \left(\mathbb{E}_{\mathbf{u}} [Y_j I_{\{\mathcal{M}(\mathbf{Y}) \geq x\}}] - v_j \mathbb{E}_{\mathbf{u}} [I_{\{\mathcal{M}(\mathbf{Y}) \geq x\}}] \right) \\ D_j''(v_j) &= w_j''(v_j) \left(\mathbb{E}_{\mathbf{u}} [Y_j I_{\{\mathcal{M}(\mathbf{Y}) \geq x\}}] - v_j \mathbb{E}_{\mathbf{u}} [I_{\{\mathcal{M}(\mathbf{Y}) \geq x\}}] \right) - w_j'(v_j) \mathbb{E}_{\mathbf{u}} [I_{\{\mathcal{M}(\mathbf{Y}) \geq x\}}]. \end{aligned}$$

It is easy to check that the definition of $w(\cdot)$ implies that $w_j'(v_j) = 1/k''(w_j(v_j)) = 1/\text{Var}[Y_j] > 0$. Therefore, at a stationary point v_j^* such that $D_j'(v_j^*) = 0$ we have

$$v_j^* = \frac{\mathbb{E}_{\mathbf{u}} [Y_j I_{\{\mathcal{M}(\mathbf{Y}) \geq x\}}]}{\mathbb{E}_{\mathbf{u}} [I_{\{\mathcal{M}(\mathbf{Y}) \geq x\}}]}}$$

(notice that the denominator is positive by the assumption $P_{\mathbf{u}}(\mathcal{M}(\mathbf{Y}) \geq x) > 0$) and, moreover,

$$D_j''(v_j^*) = -w_j'(v_j) \mathbb{E}_{\mathbf{u}} \left[I_{\{\mathcal{M}(\mathbf{Y}) \geq x\}} \right] < 0,$$

since $w_j'(v_j) = 1/\text{Var}[Y_j] > 0$. Therefore, v_j^* is a local maximum of $D_j(\cdot)$. Since v_j^* is the only point where the derivative of $D_j(\cdot)$ vanishes, it follows that v_j^* is the *unique global maximum* of $D_j(\cdot)$. ■

The vector \mathbf{v}^* in (3.4) can be estimated using the stochastic counterpart of (3.4), i.e.,

$$v_j^* \approx \frac{\sum_{i=1}^N X_{ji} I_{\{\mathcal{M}(\mathbf{X}_i) \geq x\}} W(\mathbf{X}_i, \mathbf{u}, \mathbf{v}_1)}{\sum_{i=1}^N I_{\{\mathcal{M}(\mathbf{X}_i) \geq x\}} W(\mathbf{X}_i, \mathbf{u}, \mathbf{v}_1)}, \quad (3.6)$$

where as before $\mathbf{X}_1, \dots, \mathbf{X}_N$ are random samples from the pdf $f(\mathbf{y}, \mathbf{v}_1)$.

At first sight, formulas (3.4) and (3.6) 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 those formulas are useful in terms of deriving an iterative algorithm. Also, note that for the VM programs (2.18) and (2.19) *there is no analytic solution* similar to (3.4) and (3.6) even for NEF distributions. Thus, numerical optimization procedures must be used in such case. This emphasizes one of the big advantages of the CE approach.

3.2 Finite Support Distributions

We shall classify below distributions into two categories: the ones with *unbounded support*, like exponential, Poisson and normal distributions, and the ones with *bounded support*, such as uniform $\mathcal{U}(a, b)$, truncated exponential and discrete n -point distributions. More formally, we say that $\mathcal{M}(\mathbf{Y})$ has *bounded support* if $P(|\mathcal{M}(\mathbf{Y})| > a) = 0$ for some a large enough. Note that bounded support distributions can be viewed as *zero tail* distributions as compared to their counterparts with infinite tail, which belong to the category of either *light* or *heavy* tail distributions [25]. Notice also that, in particular, finite support distributions have zero tail. As we shall see later, the optimal change of measure (the tilting distribution) in finite support distributions is a degenerated one when a is the maximum value achieved by $\mathcal{M}(\mathbf{Y})$ (see [23] for a related result). Finite support distributions play an important role in rare event probability estimation, particularly due to their connection with combinatorial optimization problems [23].

To elaborate more on finite support distributions, suppose that $\mathbf{Y}_k \sim f(\mathbf{y}, \mathbf{u})$ 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 $f(\mathbf{y}, \mathbf{v})$ with independent marginals that solves the CE problem

$$\max_{\mathbf{v}} \left\{ D(\mathbf{u}, \mathbf{v}) = \mathbb{E}_{\mathbf{u}} \left[I_{\{\mathcal{M}(\mathbf{Y}) \geq x\}} \ln f(\mathbf{Y}, \mathbf{v}) \right] : \sum_j v_{kj} = 1, k = 1, \dots, n, 0 \leq v_{kj} \leq 1 \right\}. \quad (3.7)$$

It easy to check that (3.7) 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

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

In the above, $\mathbb{E}\mathbf{u}^k$ denotes the expected value under \mathbf{u} with respect to all components except Y_k (so $\mathbf{u}^k = (u_1, \dots, u_{k-1}, u_{k+1}, \dots, u_n)$). These derivatives, together with the sufficient Karush-Kuhn-Tucker optimality conditions for problem (3.7), yield an explicit solution, which can be expressed as

$$v_{kj}^* = \frac{\mathbb{E}\mathbf{u}^k \left[I_{\{\mathcal{M}(\mathbf{Y}) \geq x\}} \mid Y_k = y_{kj} \right] u_{kj}}{\mathbb{E}\mathbf{u} \left[I_{\{\mathcal{M}(\mathbf{Y}) \geq x\}} \right]} = \frac{\mathbb{E}\mathbf{v}_1^k \left[I_{\{\mathcal{M}(\mathbf{X}) \geq x\}} W(\mathbf{X}^k, \mathbf{u}^k, \mathbf{v}_1^k) \mid X_k = y_{kj} \right] u_{kj}}{\mathbb{E}\mathbf{v}_1 \left[I_{\{\mathcal{M}(\mathbf{X}) \geq x\}} W(\mathbf{X}, \mathbf{u}, \mathbf{v}_1) \right]}, \quad (3.8)$$

provided of course that $P_{\mathbf{u}}(\mathcal{M}(\mathbf{Y}) \geq x) > 0$. In the above, \mathbf{v}_1 is an arbitrary vector of parameters, X is a random vector with distribution $f(y, \mathbf{v}_1)$, and as before \mathbf{X}^k and \mathbf{v}_1^k are respectively the vectors \mathbf{X} and \mathbf{v}_1 minus the k th component.

The vector \mathbf{v}^* in (3.8) can be estimated from the stochastic counterpart of (3.8), i.e.,

$$\begin{aligned} v_{kj}^* &\approx \frac{\sum_{i=1}^N I_{\{\mathcal{M}(X_{1i}, \dots, X_{k-1,i}, y_{kj}, X_{k+1,i}, \dots, X_{ni}) \geq x\}} W(\mathbf{X}_i^k, \mathbf{u}^k, \mathbf{v}_1^k) u_{kj}}{\sum_{j=1}^m \sum_{i=1}^N I_{\{\mathcal{M}(X_{1i}, \dots, X_{k-1,i}, y_{kj}, X_{k+1,i}, \dots, X_{ni}) \geq x\}} W(\mathbf{X}_i^k, \mathbf{u}^k, \mathbf{v}_1^k) u_{kj}} \\ &= \frac{\sum_{i=1}^N I_{\{\mathcal{M}(X_{1i}, \dots, X_{k-1,i}, y_{kj}, X_{k+1,i}, \dots, X_{ni}) \geq x\}} W(\mathbf{X}_i, \mathbf{u}, \mathbf{v}_1) (v_1)_{kj}}{\sum_{j=1}^m \sum_{i=1}^N I_{\{\mathcal{M}(X_{1i}, \dots, X_{k-1,i}, y_{kj}, X_{k+1,i}, \dots, X_{ni}) \geq x\}} W(\mathbf{X}_i, \mathbf{u}, \mathbf{v}_1) (v_1)_{kj}}, \end{aligned} \quad (3.9)$$

where $\mathbf{X}_1, \dots, \mathbf{X}_N$ are i.i.d. samples from the pmf $f(\mathbf{y}, \mathbf{v}_1)$.

3.3 The Basic CE Algorithm

Note that programs (2.20) and (2.21) (and consequently the closed formulas (3.4) and (3.8)) are useful only in the case where the original probability $\ell(x)$ defined in (2.16) is not very small, say $\ell \geq 10^{-5}$. In those cases, the CE procedure will help in terms of determining a potentially more accurate estimator. In rare-event context, however (say, $\ell(x) \leq 10^{-6}$), programs (2.20) and (2.21) do not help, since owing to the rarity of the events $\{\mathcal{M}(\mathbf{Y}_i) \geq x\}$, the random variables $I_{\{\mathcal{M}(\mathbf{Y}_i) \geq x\}}$, $i = 1, \dots, N$ and the associated derivatives of $\hat{D}_N(\mathbf{u}, \mathbf{v})$ at $\mathbf{v} = \mathbf{u}$ vanish with high probability for reasonable sizes of N .

To overcome this difficulty, we describe now a *multi-stage* algorithm. The idea is to introduce an auxiliary sequence $\{\gamma_t\}$, $t \geq 0$ and iterate in both γ_t and \mathbf{v}_t (see Algorithm 3.1 below). We start by a not very small ρ , say $\rho = 10^{-2}$. Let γ_0 ($\gamma_0 < x$) be such that, under the original pdf $f(\mathbf{y}, \mathbf{u})$, the probability $\ell(\gamma_0) = \mathbb{E}\mathbf{u} \left[I_{\{\mathcal{M}(\mathbf{Y}) \geq \gamma_0\}} \right]$ is at least ρ . We set next $\mathbf{v}_0 := \hat{\mathbf{v}}_0 := \mathbf{u}$ and then proceed iterating in both \mathbf{v} and γ with the goal of estimating the pair $\{\ell(x), \mathbf{v}^*\}$, as follows:

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

$$P_{\mathbf{v}_t}(\mathcal{M}(\mathbf{Z}) \geq \gamma_t) \geq \rho, \quad (3.10)$$

$$P_{\mathbf{v}_t}(\mathcal{M}(\mathbf{Z}) \leq \gamma_t) \geq 1 - \rho, \quad (3.11)$$

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

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

$$\hat{\gamma}_t = \hat{\gamma}_t(\mathbf{v}_t) \equiv \mathcal{M}_{(t, \lceil (1-\rho)N \rceil)}, \quad (3.12)$$

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

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

$$\max_{\mathbf{v} \in V} \left\{ D(\gamma_{t-1}, \mathbf{v}_{t-1}, \mathbf{v}) := \mathbb{E}_{\mathbf{v}_{t-1}} \left[I_{\{\mathcal{M}(\mathbf{Z}) \geq \gamma_{t-1}\}} W(\mathbf{Z}, \mathbf{u}, \mathbf{v}_{t-1}) \ln f(\mathbf{Z}, \mathbf{v}) \right] \right\}. \quad (3.13)$$

The stochastic counterpart of (3.13) is as follows: for fixed $\hat{\gamma}_{t-1}$, derive $\hat{\mathbf{v}}_t$ from the following program

$$\max_{\mathbf{v} \in V} \left\{ \hat{D}_N(\hat{\gamma}_{t-1}, \hat{\mathbf{v}}_{t-1}, \mathbf{v}) := \frac{1}{N} \sum_{j=1}^N I_{\{\mathcal{M}(\mathbf{Z}_j) \geq \hat{\gamma}_{t-1}\}} W(\mathbf{Z}_j, \mathbf{u}, \hat{\mathbf{v}}_{t-1}) \ln f(\mathbf{Z}_j, \mathbf{v}) \right\}. \quad (3.14)$$

As seen before, the optimal solutions of (3.13) and (3.14) can readily obtained *analytically*, provided $f(\mathbf{y}, \mathbf{v})$ is either a NEF or a finite support distribution — cf. (3.4), (3.6), (3.8), and (3.9). For example, the solution of (3.13) for NEF distributions is

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

whereas the solution of (3.14) 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.

The resulting algorithm for estimating $\ell(x)$ can be written as

Algorithm 3.1 :

1. Generate a sample $\mathbf{Z}_1, \dots, \mathbf{Z}_N$ from the pdf $f(\mathbf{z}, \mathbf{u})$ and compute the sample $(1 - \rho)$ -quantile (3.12). Denote the initial solution by $\hat{\gamma}_0$. Set $t := 1$.

2. Use the **same** current sample $\mathbf{Z}_1, \dots, \mathbf{Z}_N$ to solve the stochastic program (3.14). Denote the solution by $\hat{\mathbf{v}}_t := \hat{\mathbf{v}}_t(\hat{\gamma}_{t-1})$.
3. Generate a **new** sample $\mathbf{Z}_1, \dots, \mathbf{Z}_N$ from the pdf $f(\mathbf{z}, \hat{\mathbf{v}}_t)$ and compute the sample $(1 - \rho)$ -quantile (3.12). Denote the solution by $\hat{\gamma}_t$.
4. If $\hat{\gamma}_t \geq x$, set $\hat{\gamma}_t := x$ and solve the stochastic program (3.14) for $\hat{\gamma}_t = x$. Denote the solution as $\hat{\mathbf{v}}_T$ and stop; otherwise set $t := t + 1$ and reiterate from step 2.
5. Estimate the rare-event probability $\ell(x)$ using the estimate (2.17), with \mathbf{v}_1 replaced by $\hat{\mathbf{v}}_T$.
6. * (Optional) Repeat step 4 (solve the stochastic program (3.14) for $\hat{\gamma}_t = x$) for several additional iterations until desired accuracy is obtained and go to step 5.

Note that Algorithm 3.1 breaks 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 $\{(\hat{\gamma}_t, \hat{\mathbf{v}}_t)\}$ depending on the parameter (probability) ρ and such that $\rho \gg \ell(x)$.

4 Examples

For better insight into the single-stage procedure (2.20)-(2.21) and the multiple-stage procedure (3.12)-(3.14) described above, we discuss now several examples. Although in those examples the quantities of interest can be computed analytically, we present them to illustrate the algorithms. Moreover, those examples motivate the need for some assumptions, which we will have to impose when dealing with general convergence of the method.

4.1 Single-Stage Examples

Example 4.1 Suppose we are interested in estimating $\ell(x) = P(\mathcal{M}(\mathbf{Y}) \geq x)$, where $\mathcal{M}(\mathbf{Y}) = \min(Y_1, \dots, Y_n)$ and the random variables Y_1, \dots, Y_n are exponentially identically distributed with mean u , i.e., $Y_i \sim f(y, u) = 1/u \exp(-y/u)$, $i = 1, \dots, n$. Taking into account that, in this case, $\mathcal{M}(\mathbf{Y}) \geq x$ if and only if $Y_i \geq x$ for all $i = 1, \dots, n$, we have

$$\ell(x) = \prod_{i=1}^n P(Y_i \geq x) = e^{-nx/u}. \quad (4.1)$$

It follows that the SCV κ^2 (see (2.22)) for the CMC estimator is $\kappa^2(u, x) \approx e^{nx/u}$ provided x is large, hence the CMC estimator is exponential in x .

VM approach. It is shown in [25] that, for this example, the optimal value of the reference parameter $v^* = v^*(x)$ that solves (2.18) is given by

$$v^* = \left[\frac{1}{u} + \frac{1}{x} - \sqrt{\frac{1}{u^2} + \frac{1}{x^2}} \right]^{-1}. \quad (4.2)$$

Suppose $x \gg u$, hence $P(\mathcal{M}(\mathbf{Y}) \geq x)$ is small, say less than 10^{-6} . In this case we have $v^* \approx x$ and thus, from (4.1) we get $\mathbb{E}_{v^*} [I_{\{\mathcal{M}(\mathbf{Y}) \geq x\}}] \approx e^{-n}$.

Let us compute the SCV $\kappa^2(v, x)$ of the LR estimator $\hat{\ell}_N(x, v)$ in (2.17). It is easy to see that, since Y_1, \dots, Y_n are i.i.d. and exponentially distributed, we have

$$\kappa^2(v, x) = \left(\frac{v^2}{u(2v-u)} \right)^n \frac{P_{uv/(2v-u)}(\mathcal{M}(\mathbf{Y}) \geq x)}{(P_u(\mathcal{M}(\mathbf{Y}) \geq x))^2} - 1, \quad (4.3)$$

so for the particular function \mathcal{M} above we obtain

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

For $v^* = x$ it reduces to $\kappa^2(v^*, x) \approx x^n e^n / (2u)^n$. That is, for large x , the SCV of the CMC increase in x exponentially, whereas the SCV of the optimal LR estimator increase in x polynomially.

In other words, the CMC and VM estimators can be viewed as having respectively *exponential* and *polynomial* complexity. For a more general discussion on complexity, see [2].

CE approach. The formulas for the CE approach can be easily derived by noticing that, since the exponential distribution belongs to the NEF family, we can apply formula (3.4) directly. It follows that

$$v^* = x + u. \quad (4.4)$$

For $x \gg u$ it gives $v^* \approx x$, which coincides with the asymptotic value obtained for the VM approach. Clearly, $\kappa^2(v^*, x)$ for both VM and CE coincide.

Example 4.2 Suppose we are interested in estimating $\ell(x) = P(\mathcal{M}(\mathbf{Y}) \geq x)$, where $\mathcal{M}(\mathbf{Y}) = Y_1 + \dots + Y_n$ and, as in Example 4.1, the random variables Y_1, \dots, Y_n are exponentially identically distributed with mean u . Of course, in that case we know that $\mathcal{M}(\mathbf{Y})$ has Gamma distribution with parameters n and $1/u$, so $P(\mathcal{M}(\mathbf{Y}) \geq x)$ can be computed exactly as

$$P(\mathcal{M}(\mathbf{Y}) \geq x) = \sum_{k=0}^{n-1} \frac{e^{-x/u} (x/u)^k}{k!}.$$

It is difficult to compute the VM-optimal parameter in this case. However, we can compute the CE-optimal parameter \mathbf{v}^* given by (3.4). Since Y_1, \dots, Y_n are i.i.d., it is clear that all components of \mathbf{v}^* are identical. After some algebra, we obtain that

$$v^* = \frac{u \sum_{k=0}^n e^{-x/u} (x/u)^k / k!}{\sum_{k=0}^{n-1} e^{-x/u} (x/u)^k / k!} = \frac{u P(Y_1 + \dots + Y_{n+1} \geq x)}{P(Y_1 + \dots + Y_n \geq x)}, \quad (4.5)$$

where Y_{n+1} is independent of Y_1, \dots, Y_n and has the same distribution as Y_1, \dots, Y_n . From the above expression, it is easy to see that, when x/u is large (i.e. $x \gg u$), we have

$$v^* \approx u(1 + x/u) = u + x \approx x.$$

Let us compute the SCV $\kappa^2(v, x)$ of the LR estimator $\widehat{\ell}_N(x, v)$. Using exact values in (4.3), however, does not help in terms of understanding the asymptotic behavior of κ as $x \rightarrow \infty$. Instead, we shall use a *large deviations* approximation of the probabilities in (4.3). Let $\alpha := x/(nu)$. Then we have, for sufficiently large n (and $x > nu$)

$$P_u(Y_1 + \dots + Y_n \geq x) = P_u([Y_1 + \dots + Y_n]/n \geq \alpha u) \approx C_n(u, x)e^{-nI(\alpha u)},$$

where $I(z)$ is the *rate function* of the exponential distribution, which is given by $I(z) = z/u - 1 - \ln(z/u)$ (see, e.g., [6]), and $C_n(u, x)$ is given by

$$C_n(u, x) = \frac{\text{Var}[Y_1]}{(x/n)\sqrt{2\pi n}} = \frac{u}{\alpha\sqrt{2\pi n}} = \frac{u^2}{x} \sqrt{\frac{n}{2\pi}}. \quad (4.6)$$

It follows that $e^{-nI(\alpha u)} = (\alpha e^{-(\alpha-1)})^n$ and thus

$$P_u(Y_1 + \dots + Y_n \geq x) \approx C_n(u, x) \left(\frac{x}{nu} e^{1-x/(nu)} \right)^n. \quad (4.7)$$

By substituting these values into (4.3), we get

$$\kappa^2(v, x) \approx n^n \left(\frac{v}{x} \right)^n e^{x/v-n} \frac{v^2 x}{u^2(2v-u)^2}$$

and hence, for $v^* = x$ we have

$$\kappa^2(v^*, x) \approx n^n e^{1-n} \frac{x^3}{u^2(2x-u)^2}.$$

That is, the SCV of the optimal LR estimator grows *linearly* with respect to x . Notice that, for $v = u$ — which corresponds to the CMC estimator — $\kappa^2(v, x)$ increase in x exponentially.

Example 4.3 Suppose we are again interested in estimating $\ell(x) = P(\mathcal{M}(\mathbf{Y}) \geq x)$, where $\mathcal{M}(\mathbf{Y}) = \min(Y_1, \dots, Y_n)$, but now the random variables Y_1, \dots, Y_n are i.i.d. and each has discrete distribution with $P(Y_i = j) = u_j$, $j = 1, \dots, m$. For simplicity of notation, suppose that x is integer. Again, since $\mathcal{M}(\mathbf{Y}) \geq x$ if and only $Y_i \geq x$ for all $i = 1, \dots, n$, we have

$$\ell(x) = \prod_{i=1}^n P(Y_i \geq x) = \left(\sum_{j=x}^m u_j \right)^n.$$

The VM function $\mathcal{V}(\mathbf{u}, \mathbf{v})$ in (2.18) reduces to

$$\mathcal{V}(\mathbf{u}, \mathbf{v}) = \mathbb{E}_{\mathbf{v}} \left[I_{\{\mathcal{M}(\mathbf{Z}) \geq x\}} W^2(\mathbf{Z}, \mathbf{u}, \mathbf{v}) \right] = \left(\sum_{j=1}^m I_{\{j \geq x\}} \frac{u_j^2}{v_j} \right)^n,$$

whereas the CE function $D(\mathbf{u}, \mathbf{v})$ in (2.20) can be written as

$$D(\mathbf{u}, \mathbf{v}) = \left(\sum_{j=x}^m u_j \ln v_j \right)^n .$$

Minimizing $\mathcal{V}(\mathbf{u}, \mathbf{v})$ (or maximizing $D(\mathbf{u}, \mathbf{v})$) over \mathbf{v} under the condition $\sum_{i=1}^n v_i = 1$ gives in both cases

$$v_i^* = \begin{cases} 0 & \text{if } 1 \leq i \leq x-1 \\ \frac{u_i}{\sum_{j=x}^m u_j} & \text{if } x \leq i \leq m . \end{cases} \quad (4.8)$$

Note that for $x = m$ (4.8) simplifies to

$$v_i^* = \begin{cases} 0 & 1 \leq i \leq m-1 \\ 1 & i = m. \end{cases} \quad (4.9)$$

Taking into the account that $\ell(m) = u_m^n$, it is readily seen that, for $\mathbf{v} = \mathbf{v}^*$, the variance $\text{Var}[\widehat{\ell}_N(x, \mathbf{v}^*)]$ of the LR estimator given in (2.17) is *zero*. Clearly, in this case the SCV $\kappa^2(\mathbf{v}^*, x) = 0$.

The above example illustrates the following important general property of *finite support* distributions:

Proposition 4.1 *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 \mathbf{y}^) is unique. Suppose that the random vector \mathbf{Y} has independent components with discrete distribution on \mathcal{Y} . Then, the solution of both VM and CE programs (2.18) and (2.20) for $P(\mathcal{M}(\mathbf{Y}) \geq x^*)$ is the atomic measure (we shall also call it degenerate) with mass at \mathbf{y}^* .*

Proof. Let \mathbf{v}_d^* denote the degenerate measure with mass on \mathbf{y}^* . That \mathbf{v}_d^* solves (2.18) follows immediately from the fact that the variance of estimator $\widehat{\ell}_N(x^*)$ given in (2.17), under \mathbf{v}_d^* , is *zero*.

Let $f(\mathbf{y}, \mathbf{u})$ denote the distribution of \mathbf{Y} . Consider now formula (3.8), derived for finite support distributions. Notice that the term $\mathbb{E}\mathbf{u} \left[I_{\{\mathcal{M}(\mathbf{Y}) \geq x^*\}} | Y_i = y_{ij} \right]$ is equal to zero if $y_{ij} \neq y_i^*$. Otherwise, we have

$$\mathbb{E}\mathbf{u} \left[I_{\{\mathcal{M}(\mathbf{Y}) \geq x^*\}} | Y_i = y_i^* \right] = \mathbb{E}\mathbf{u} \left[I_{\{\mathbf{Y} = \mathbf{y}^*\}} | Y_i = y_i^* \right] = \prod_{k \neq i} P\mathbf{u}(Y_k = y_k^*),$$

and so in (3.8) we obtain that

$$v_{ij}^* = \begin{cases} 0 & \text{if } y_{ij} \neq y_i^*, \text{ and} \\ \frac{\prod_{k \neq i} P\mathbf{u}(Y_k = y_k^*) u_{ij}}{\prod_k P\mathbf{u}(Y_k = y_k^*)} = 1 & \text{otherwise.} \end{cases}$$

■

Proposition 4.1 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 $P(\mathcal{M}(\mathbf{Y}) \geq x)$ are always the same, *regardless of the distribution of \mathbf{Y}* . This property in turn has nice implications for combinatorial optimization; see [23] for a discussion.

It is also worth mentioning that the assumption of uniqueness of the maximizer of \mathcal{M} in Proposition 4.1 can be artificially enforced by imposing some ordering on the finite set \mathcal{Y} , say the lexicographical order. Then, we can define a function \mathcal{Z} on \mathcal{Y} as $\mathcal{Z}(\mathbf{y}) = \mathcal{M}(\mathbf{y}) - \varepsilon(\mathbf{y})$, where $\varepsilon(\mathbf{y})$ is a small perturbation, proportional to the rank of \mathbf{y} and small enough to ensure that $\mathcal{Z}(\mathbf{y}^1) > \mathcal{Z}(\mathbf{y}^2)$ if $\mathcal{M}(\mathbf{y}^1) > \mathcal{M}(\mathbf{y}^2)$. In that case, the degenerate measure v^* has mass at the element with highest rank within the set of maximizers of \mathcal{M} .

4.2 Multi-Stage Examples

We extend now the single-stage examples of section 4.1 to the corresponding multiple stages ones to illustrate the behavior of Algorithm 3.1. For simplicity, we shall use the deterministic version (3.10), (3.11) and (3.13) instead of their stochastic counterparts (3.12) and (3.14).

Example 4.4 Example 4.1 continued. Recall that in that example we had $\mathcal{M}(\mathbf{Y}) = \min(Y_1, \dots, Y_n)$ with $Y_i \sim f(y, u) = 1/ue^{-y/u}$. As seen earlier, the CE-optimal solution is $v^* = x + u$ (cf. (4.4)). Consider now γ_t defined in (3.10)-(3.11). Since the algorithm stops when $\gamma_t \geq x$, and since the distribution of $\mathcal{M}(\mathbf{Y})$ is continuous, we can write

$$\gamma_t = \max \{ \gamma \leq x : \exp(-\gamma n/v_t) \geq \rho \} = \min \{ x, Cv_t/n \},$$

where $C = \log(1/\rho) > 0$. The parameter v_t defined in (3.13) can then be rewritten using (3.4) as

$$v_{t+1} = \gamma_t + u = \min \{ x, Cv_t/n \} + u. \quad (4.10)$$

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} . From (4.4) and (4.10) we can see that $\bar{v} = v^*$ if and only if

$$x \leq (C/n)\bar{v}. \quad (4.11)$$

Let us compute \bar{v} . If $C/n \geq 1$, i.e., if $C \geq n$, then we have $\bar{v} = x + u = v^*$. On the other hand, if $C/n < 1$, i.e., if $C < n$, then we have two cases:

1. $x \leq (C/n)(x + u)$: in this case, we have that $g(x + u) = x + u$, i.e. $\bar{v} = x + u$.
2. $x > (C/n)(x + u)$: in this case, we have that $\bar{v} = u/(1 - C/n)$.

In any of these events, it is easy to check that the optimality condition (4.11) becomes $x \leq (C/n)(x + u)$, i.e. $C \geq nx/(x + u)$. Since $C = \log(1/\rho)$, it follows that Algorithm 3.1 will converge to the correct solution if and only if

$$\rho \leq \exp\left(-\frac{nx}{x+u}\right). \quad (4.12)$$

Moreover, if $\rho \leq \exp(-n)$ (which implies (4.12)), 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 ; if $\rho > \exp(-n)$, then the differences $v_{t+1} - v_t$ *decrease* until the point when x is hit by γ_t .

At first sight, condition (4.12) seems discouraging, since it requires the parameter ρ to decrease exponentially with n . Notice however that this example constitutes an intrinsically difficult problem — from (4.1), we see that the probability being estimated goes to zero exponentially in n under *any* parameter. It makes perhaps more sense to consider the behavior of (4.12) for *fixed* n — then we see that $\rho \leq \exp(-n)$ is a sufficient condition for Algorithm 3.1 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 (4.12) becomes asymptotically $\rho \leq \exp(-\Delta/u)$.

Example 4.5 Example 4.2 continued. Recall that in that example we had $\mathcal{M}(\mathbf{Y}) = Y_1 + \dots + Y_n$ with $Y_i \sim f(y, u) = 1/ue^{-y/u}$. As seen earlier, the CE-optimal solution is given by (4.5). In order to conduct an analytical study, in what follows we will assume that $x = \alpha nu$ for some $\alpha > 1$. Moreover, we shall make extensive use of the large deviations approximation (4.7) for $P_u(Y_1 + \dots + Y_n \geq x)$, which is valid when n is large and $x > nu$. To simplify the calculations, however, we shall write the term $C_n(u, x)$ in (4.6) as $C_n(u, x) \approx K/\sqrt{n}$, where K is a positive constant. Notice also that (4.7) implies that $P(\mathcal{M}(\mathbf{Y}) \geq x)$ goes to zero exponentially fast on n when $x = \alpha nu$.

Consider initially γ_t defined in (3.10)-(3.11). Assuming $\gamma_t > nv_t$, we can determine γ_t by solving the inequality

$$\frac{K}{\sqrt{n}} \left(\frac{\gamma_t}{nv_t} e^{1-\gamma_t/(nv_t)} \right)^n \geq \rho.$$

By applying logarithm on both sides and defining $\alpha_t := \gamma_t/(nv_t) > 1$, we obtain

$$\ln \alpha_t + (1 - \alpha_t) \geq \ln[(\rho\sqrt{n}/K)^{1/n}].$$

Let $\theta := \ln[(\rho\sqrt{n}/K)^{1/n}]$. Using the approximation $\ln \alpha_t \approx (\alpha_t - 1) - (\alpha_t - 1)^2/2$ and solving the above inequality, we have $\alpha_t \leq 1 + \sqrt{-2\theta}$ (provided $\theta < 0$) and thus $\gamma_t \leq nv_t(1 + \sqrt{-2\theta})$, so that

$$\gamma_t = \min\{x, nv_t(1 + \sqrt{-2\theta})\}.$$

Let us compute v_{t+1} . Using (4.7) in formula (4.5) (with γ_t in place of x), we obtain

$$v_{t+1} \approx \frac{e\gamma_t}{n} \left(\frac{n}{n+1}\right)^{n+3/2} = \min\left\{\frac{ex}{n}, ev_t(1 + \sqrt{-2\theta})\right\} \left(\frac{n}{n+1}\right)^{n+3/2}.$$

It follows that, if the coefficient of v_t is bigger than one, then the sequence $\{v_t\}$ increases exponentially until the boundary given by the min operation above is achieved. This occurs if and only if $1 + \sqrt{-2\theta} > (1 + n^{-1})^{n+3/2} e^{-1}$. Since $(1 + 1/n)^n \approx e$, the above inequality becomes $1 + \sqrt{-2\theta} > (1 + 1/n)^{3/2}$. By imposing that $1 + \sqrt{-2\theta} > (1 + 1/n)^2$ we obtain

$$\rho < \frac{K}{\sqrt{n}} e^{-\frac{1}{2n^3} + \frac{1}{n^2} - \frac{2}{n}}.$$

Though the approximations used in the above calculations preclude the computation of an exact condition for ρ , we can see that ρ must decrease at rate $n^{-1/2}$, which constitutes a mild requirement.

Example 4.6 Example 4.3 continued. For simplicity of notation, suppose that $u_j = 1/m$ for all j . It is readily seen that in this case (4.8) becomes

$$v_i^* = \begin{cases} 0 & 1 \leq i \leq x - 1 \\ \frac{1}{m - x + 1} & x \leq i \leq m. \end{cases} \quad (4.13)$$

By definition (3.10)-(3.11), γ_0 must satisfy $\left(\frac{m-\gamma_0+1}{m}\right)^n \geq \rho$ and $\left(\frac{\gamma_0}{m}\right)^n \geq 1 - \rho$. Suppose we pick the *largest* γ_0 satisfying the above properties. Then, $\gamma_0 = 1 + \lfloor m(1 - \rho^{1/n}) \rfloor$.

Let us now compute γ_{t+1} as a function of γ_t . From (4.13) (with γ_t in place of x) we have that

$$v_{ti} = \begin{cases} 0 & 1 \leq i \leq \gamma_t - 1 \\ \frac{1}{m - \gamma_t + 1} & \gamma_t \leq i \leq m. \end{cases}$$

It follows (again using the largest quantile criterion) that that we can compute γ_{t+1} as the largest solution of

$$\left(\frac{m - \gamma + 1}{m - \gamma_t + 1}\right)^n \geq \rho,$$

that is,

$$\gamma_{t+1} = \lfloor (m + 1)(1 - \rho^{1/n}) + \rho^{1/n} \gamma_t \rfloor. \quad (4.14)$$

Due to the rounding operation, it is difficult to compute a closed form solution for γ_t . We can however use the bounds

$$(m+1)(1-\rho^{1/n})+\rho^{1/n}\gamma_t-1 \leq \gamma_{t+1} \leq (m+1)(1-\rho^{1/n})+\rho^{1/n}\gamma_t.$$

By applying these bounds recursively, we obtain that

$$\begin{aligned} \gamma_{t+1} &\geq \left((m+1)(1-\rho^{1/n})-1\right)(1+\rho^{1/n}+\dots+\rho^{(t+1)/n}) \\ \gamma_{t+1} &\leq \left((m+1)(1-\rho^{1/n})\right)(1+\rho^{1/n}+\dots+\rho^{(t+1)/n}), \end{aligned}$$

i.e.

$$\left(m+1-\frac{1}{1-\rho^{1/n}}\right)(1-\rho^{(t+1)/n}) \leq \gamma_{t+1} \leq (m+1)(1-\rho^{(t+1)/n}).$$

We can infer from the above inequality that, if $m+1-\frac{1}{1-\rho^{1/n}}>x-1$, i.e. if $\rho<\left(\frac{m-x+1}{m-x+2}\right)^n$, then for t large enough we have $\gamma_{t+1}\geq x$. Therefore, at some stage $r+1$ of Algorithm 3.1, after Step 2 we obtain that $\gamma_{t+1}=x$ (unless $x>m$, in which case $P(\mathcal{M}(\mathbf{Y})\geq x)=0$). It follows that, after steps 3 and 4 of Algorithm 3.1, we obtain \mathbf{v}_{r+1} equal to \mathbf{v}^* in (4.13). In other words, the algorithm finishes after a *finite* number of steps provided

$$\rho < \left(\frac{m-x+1}{m-x+2}\right)^n. \quad (4.15)$$

As in Example 4.4, one should not be discouraged by the fact that ρ must decrease exponentially with n ; rather, one should notice that, since $(m-x+1)/(m-x+2)\geq 1/2$ for $x\leq m$, it follows that $\rho<2^{-n}$ suffices for Algorithm 3.1 to work whenever x belongs to the interval $[0,m]$. Moreover, condition (4.15) is quite conservative — indeed, (4.15) requires an exponential bound on ρ even when $x=0$.

5 Convergence of the CE Method

We discuss now some issues related to convergence of the CE method. Notice initially that, as the examples in section 4 illustrate, 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 shall propose a modification of the basic CE algorithm where ρ is changed *adaptively* (see [23] for related ideas). For $\mathbf{z}\in\mathbb{R}^n$, $\mathbf{v}\in\mathbb{R}^m$, and $\rho>0$, define $\gamma(\mathbf{v},\rho)$ as an arbitrary $(1-\rho)$ -quantile of $\mathcal{M}(\mathbf{Z})$ under \mathbf{v} (cf. (3.10)-(3.11)). Let

$$\psi(\mathbf{z},\mathbf{v},\rho) := I_{\{\mathcal{M}(\mathbf{z})\geq\min(x,\gamma(\mathbf{v},\rho))\}}. \quad (5.1)$$

Let \mathbf{v}^* be a CE-optimal solution, i.e. a maximizer of $D(\mathbf{u}, \mathbf{v})$ defined in (2.20) with respect to \mathbf{v} . That is, we have that

$$\mathbf{v}^* \in \operatorname{argmax}_{\mathbf{v} \in V} \left\{ \mathbb{E}_{\mathbf{u}} \left[I_{\{\mathcal{M}(\mathbf{Y}) \geq x\}} \ln f(\mathbf{Y}, \mathbf{v}) \right] \right\}. \quad (5.2)$$

We will need the following assumption:

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

Assumption A simply ensures that the probability being estimated — $P_{\mathbf{u}}(\mathcal{M}(\mathbf{Z}) \geq x)$ — does not vanish when \mathbf{u} is replaced by a feasible parameter $\mathbf{v} \in V$. The assumption is trivially satisfied when the distribution of $\mathcal{M}(\mathbf{Z})$ has infinite tail. For zero tail distributions (cf. section 3.2), the assumption holds as long as either x is less than the maximum value of the function $\mathcal{M}(\mathbf{Z})$, or if there is a positive probability that x is attained.

Let $\rho_x^* := P_{\mathbf{v}^*}(\mathcal{M}(\mathbf{Z}) \geq x)$. By assumption A, $\rho_x^* > 0$. Let $\rho^* \in (0, \rho_x^*)$ be arbitrary. By the definition of γ , we have that

$$\begin{aligned} P_{\mathbf{v}^*}(\mathcal{M}(\mathbf{Z}) \geq \gamma(\mathbf{v}^*, \rho^*)) &\geq \rho^* \\ P_{\mathbf{v}^*}(\mathcal{M}(\mathbf{Z}) \leq \gamma(\mathbf{v}^*, \rho^*)) &\geq 1 - \rho^* > 1 - \rho_x^*. \end{aligned} \quad (5.3)$$

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

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

which contradicts (5.3). It follows that $\gamma(\mathbf{v}^*, \rho^*) \geq x$ and thus $\psi(\mathbf{Z}, \mathbf{v}^*, \rho^*) = I_{\{\mathcal{M}(\mathbf{Z}) \geq x\}}$. We can then re-write the function $D(\mathbf{u}, \mathbf{v})$ defined in (2.20) (with $\mathbf{v}_1 = \mathbf{u}$) as

$$\begin{aligned} D(\mathbf{u}, \mathbf{v}) &= Q(\mathbf{v}, \mathbf{v}^*, \rho^*) := \mathbb{E}_{\mathbf{u}} [\psi(\mathbf{Y}, \mathbf{v}^*, \rho^*) \ln f(\mathbf{Y}, \mathbf{v})] \\ &= \mathbb{E}_{\mathbf{v}^*} [\psi(\mathbf{Z}, \mathbf{v}^*, \rho^*) W(\mathbf{Z}, \mathbf{u}, \mathbf{v}^*) \ln f(\mathbf{Z}, \mathbf{v})]. \end{aligned}$$

The new algorithm — in its deterministic version — is stated as follows:

Algorithm 5.1 :

0. Let $t := 1$, $\rho_0 := \rho$, $\mathbf{v}_0 := \mathbf{u}$.

1. Compute

$$Q(\mathbf{v}, \mathbf{v}_{t-1}, \rho_{t-1}) = \mathbb{E}_{\mathbf{v}_{t-1}} [\psi(\mathbf{Z}, \mathbf{v}_{t-1}, \rho_{t-1}) W(\mathbf{Z}, \mathbf{u}, \mathbf{v}_{t-1}) \ln f(\mathbf{Z}, \mathbf{v})].$$

2. Compute

$$\mathbf{v}_t \in \operatorname{argmax}_{\mathbf{v} \in V} \{Q(\mathbf{v}, \mathbf{v}_{t-1}, \rho_{t-1})\}.$$

3. If $\gamma(\mathbf{v}_{t-1}, \rho_{t-1}) \geq x$, STOP.

4. Otherwise, let ρ_t be such that $\gamma(\mathbf{v}_t, \rho_t) \geq \min\{x, \gamma(\mathbf{v}_{t-1}, \rho_{t-1}) + \delta\}$, where δ is a positive constant.

5. Let $t := t + 1$ and go back to step 1.

Notice that without steps 3 and 4 (and with $\rho_t \equiv \rho$), the above algorithm is just a re-statement of Algorithm 3.1. Furthermore, in the above form the CE algorithm can be viewed in the same spirit as the well known *Expectation-Maximization* (EM) algorithm, widely known in the statistics literature. Generally speaking, the purpose of the EM algorithm is to compute maximum likelihood estimates when part of the data is missing (see, e.g., [21]). Without steps 3 and 4 above, Algorithm 5.1 has exactly the form of an EM method — step 1 is the expectation (E) step, whereas step 2 is the maximization (M) step. Algorithm 5.1, however, is more complex than the original EM algorithm in that it involves the ψ and W functions in addition to the log-likelihood term.

With the proposed modifications — i.e. steps 3 and 4 — one can verify that Algorithm 5.1 converges to \mathbf{v}^* in *finite* number of steps. This is formalized in the proposition below.

Proposition 5.1 *Suppose that Assumption A holds. Then, Algorithm 5.1 converges to a CE-optimal solution after a finite number of iterations.*

Proof. Let t be an arbitrary iteration of the algorithm. Let $\rho_x := P_{\mathbf{v}_t}(\mathcal{M}(\mathbf{Z}) \geq x)$. By assumption A, $\rho_x > 0$, since $\mathbf{v}_t \in V$. Let $\rho^* \in (0, \rho_x)$ be arbitrary. Using similar arguments as in the discussion following (5.3), one can show that $\gamma(\mathbf{v}_t, \rho^*) \geq x$. Thus, step 4 of Algorithm 5.1 can always be accomplished. It follows that the algorithm stops after T iterations, with $T \leq \lceil x/\delta \rceil$. At that point, we have $\gamma(\mathbf{v}_{T-1}, \rho_{T-1}) \geq x$. Next, observe that step 2 of Algorithm 5.1 can be re-written as

$$\mathbf{v}_t \in \operatorname{argmax}_{\mathbf{v} \in V} \left\{ \mathbb{E}_{\mathbf{u}} \left[I_{\{\mathcal{M}(\mathbf{Y}) \geq \min\{x, \gamma(\mathbf{v}_{t-1}, \rho_{t-1})\}\}} \ln f(\mathbf{Y}, \mathbf{v}) \right] \right\}.$$

Thus, at $t = T$ we have

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

From (5.2), we see that \mathbf{v}_T is a CE-optimal solution. Thus, Algorithm 5.1 converges to the CE-optimal solution in finite number of iterations. ■

A few comments about Algorithm 5.1 are now in order. First, notice that the algorithm aims to reach \mathbf{v}^* by a sequence of calculations controlled by the parameters ρ_t , which should be kept as large as possible. In fact, in many cases ρ_t is initialized with a not very small number, say, 0.01, and is never reduced (see, e.g., [23, 24]). This follows from the fact that, by construction of the CE problem, the distribution of $\mathcal{M}(\mathbf{Z})$ when $\mathbf{Z} \sim f(\mathbf{z}, \mathbf{v}_{t+1})$ tends to have larger tails than when $\mathbf{Z} \sim f(\mathbf{z}, \mathbf{v}_t)$ and thus $\gamma(\mathbf{v}_{t+1}, \rho) > \gamma(\mathbf{v}_t, \rho)$ for all ρ in those cases.

In that sense, step 4 of Algorithm 5.1 can be viewed as a “safeguard” for the cases when the latter property fails. As discussed above, there always exists ρ_{t+1} satisfying the property in step 4 of the algorithm. In practice, one can often use $\delta = 0$ in the algorithm and still obtain a convergent algorithm (see, e.g., [23, 24]). There are other alternatives as well — one can keep, for example, ρ_t constant until it is detected that the differences $\gamma(\mathbf{v}_t, \rho_t) - \gamma(\mathbf{v}_{t-1}, \rho_{t-1})$ are starting to decrease, at which point ρ_t is reduced. We will comment more on that later.

Another aspect of Algorithm 5.1 to be emphasized is that the algorithm described above is *conceptual*, in the sense that it assumes that expected values can be computed. In practice, of course, Algorithm 5.1 will be used with sample averages replacing expected values — indeed, if expected values could be computed, then one could determine \mathbf{v}^* directly from (5.2). Thus, step 2 of Algorithm 5.1 is written as

$$\hat{\mathbf{v}}_t \in \operatorname{argmax}_{\mathbf{v} \in V} \left\{ \frac{1}{N} \sum_{i=1}^N I_{\{\mathcal{M}(\mathbf{Z}_i) \geq \hat{\gamma}(\mathbf{Z}, \rho_{t-1})\}} W(\mathbf{Z}_i, \mathbf{u}, \hat{\mathbf{v}}_{t-1}) \ln f(\mathbf{Z}_i, \mathbf{v}) \right\},$$

where $\mathbf{Z}_1, \dots, \mathbf{Z}_N$ is a sample from $f(\mathbf{z}, \hat{\mathbf{v}}_{t-1})$ (for a given realization of $\hat{\mathbf{v}}_{t-1}$) and $\hat{\gamma}_N(\mathbf{Z}, \rho_{t-1})$ is defined as the sample $(1 - \rho_{t-1})$ -quantile of $\mathcal{M}(\mathbf{Z}_1), \dots, \mathcal{M}(\mathbf{Z}_N)$.

The fact that $\gamma(\mathbf{v}, \rho_{t-1})$ is replaced by $\hat{\gamma}_N(\mathbf{Z}, \rho_{t-1})$ means that, in principle, step 4 of Algorithm 5.1 may not be accomplishable. For example, it could happen that the samples $\mathcal{M}(\mathbf{Z}_i)$ produced at two consecutive iterations are identical. The proposition below shows that, when N is large enough, step 4 can be executed. In the proposition, the term “with probability one” refers to the probability space where \mathbf{Z} lies, and when $\mathbf{Z}_1, \mathbf{Z}_2, \dots$ are viewed as random variables on that space.

Proposition 5.2 *Suppose assumption A holds. Let $\mathbf{v} \in V$, and let $\mathbf{Z}_1, \mathbf{Z}_2, \dots$ be i.i.d. with common density $f(\mathbf{z}, \mathbf{v})$. Then, there exists $\rho_x > 0$ and a random $N_x > 0$ such that, with probability one, $\hat{\gamma}_N(\mathbf{Z}, \rho) \geq x$ for all $\rho \in (0, \rho_x)$ and all $N \geq N_x$. Moreover, the probability that $\hat{\gamma}_N(\mathbf{Z}, \rho) \geq x$ for a given N goes to one 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 $\partial_{\xi}\mathbb{E}\phi(Y, \xi)$ can be expressed as

$$\partial_{\xi}\mathbb{E}\phi(Y, \xi) = [\rho - P(Y \geq \xi), -(1 - \rho) + P(Y \leq \xi)].$$

Therefore, ξ satisfies the optimality condition $0 \in \partial\mathbb{E}\phi(Y, \xi)$ if and only if

$$\begin{aligned} \rho - P(Y \geq \xi) &\leq 0 \\ -(1 - \rho) + P(Y \leq \xi) &\geq 0 \end{aligned}$$

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, \dots, Y_N (call it $\hat{\xi}_{\rho}$) 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 between $\hat{\xi}_{\rho}$ and the set of $(1 - \rho)$ -quantiles of Y goes to zero as N goes to infinity w.p.1 (see, e.g., [26]).

Let $\mathbf{Z}_1, \dots, \mathbf{Z}_N$ be i.i.d. samples from $f(\mathbf{z}, \mathbf{v})$. Consider initially the case where $P_{\mathbf{v}}(\mathcal{M}(\mathbf{Z}) > x) > 0$. Using again similar arguments as in the discussion following (5.3), we obtain that $\gamma(\mathbf{v}, \rho^*) > x$ for any $\rho^* \in (0, \rho_x^+)$, where $\rho_x^+ = P_{\mathbf{v}}(\mathcal{M}(\mathbf{Z}) > x) > 0$. As discussed above, the distance between the sample $(1 - \rho^*)$ -quantile $\hat{\gamma}_N(\mathbf{Z}, \rho^*)$ of $\mathcal{M}(\mathbf{Z}_1), \dots, \mathcal{M}(\mathbf{Z}_N)$ and the set of $(1 - \rho^*)$ -quantiles of $\mathcal{M}(\mathbf{Z})$ goes to zero as N goes to infinity w.p.1. Since $\gamma(\mathbf{v}, \rho^*) > x$, it follows that $\hat{\gamma}_N(\mathbf{Z}, \rho^*) > x$ w.p.1 for N large enough. Moreover, the probability that $\hat{\gamma}_N(\mathbf{Z}, \rho^*) > x$ for a given N goes to one exponentially fast (see, e.g., [17]).

Consider now the case where $P_{\mathbf{v}}(\mathcal{M}(\mathbf{Z}) > x) = 0$, i.e. x is the maximum value achieved by $\mathcal{M}(\mathbf{Z})$. By assumption A, this implies that $\rho_x^0 := P_{\mathbf{v}}(\mathcal{M}(\mathbf{Z}) = x) > 0$ and thus, for any $\rho^* \in (0, \rho_x^0)$ we must have $\gamma(\mathbf{v}, \rho^*) = x$. It follows that $\gamma(\mathbf{v}, \rho^*) = x$ is also the unique $(1 - \rho^*)$ -quantile of the random variable $W := xI_{\{\mathcal{M}(\mathbf{Z})=x\}}$. It is clear that $\hat{\gamma}_N^x := xI_{\{\hat{\gamma}_N(\mathbf{Z}, \rho^*)=x\}}$ is a sample $(1 - \rho^*)$ -quantile of W_1, \dots, W_N , where $W_i := xI_{\{\mathcal{M}(\mathbf{Z}_i)=x\}}$. Since the distribution of W has finite support, it follows from the results in [28] that $\hat{\gamma}_N^x = \gamma(\mathbf{v}, \rho^*) = x$ w.p.1 for N large enough, and, moreover, the probability that $\hat{\gamma}_N^x = \gamma(\mathbf{v}, \rho^*) = x$ for a given N goes to one exponentially fast. Since $\hat{\gamma}_N^x = x$ if and only if $\hat{\gamma}_N(\mathbf{Z}, \rho^*) = x$, the proof is complete. \blacksquare

The above proposition shows not only that $\hat{\gamma}_N(\mathbf{Z}, \rho)$ reaches 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). In any case, in order to ensure that the sample size grows as needed, we can check, after step 4 of Algorithm 5.1, whether such ρ_{t+1} could be found. If not, then the sample size is increased and the process repeated.

Motivated by the above considerations, we present below a modified version of Algorithm 3.1. The algorithm requires the definition of constants ρ (typically, $0.01 \leq \rho \leq 0.1$), $\alpha > 1$ and $\delta > 0$.

Algorithm 5.2 :

1. Set $\rho_0 := \rho$, $N :=$ initial sample size. Generate a sample $\mathbf{Z}_1, \dots, \mathbf{Z}_N$ from the pdf $f(\mathbf{z}, \mathbf{u})$ and compute the sample $(1 - \rho_0)$ -quantile (3.12). Denote the initial solution by $\hat{\gamma}_0$. Set $t := 1$.
2. Use the **same** current sample $\mathbf{Z}_1, \dots, \mathbf{Z}_N$ to solve the stochastic program (3.14). Denote the solution by $\hat{\mathbf{v}}_t := \hat{\mathbf{v}}_t(\hat{\gamma}_{t-1})$.
3. Generate a **new** sample $\mathbf{Z}_1, \dots, \mathbf{Z}_N$ from the pdf $f(\mathbf{z}, \hat{\mathbf{v}}_t)$. Let $\rho_t := \rho$.
4. Compute the sample $(1 - \rho_t)$ -quantile (3.12). Denote the solution by $\hat{\gamma}_t$.
5. If $\hat{\gamma}_t \geq x$, set $\hat{\gamma}_t := x$ and solve the stochastic program (3.14) for $\hat{\gamma}_t = x$. Denote the solution as $\hat{\mathbf{v}}_T$ and go to step 7.
6. Otherwise, check whether there exists $\bar{\rho}$ such that the sample $(1 - \bar{\rho})$ -quantile of $\mathcal{M}(\mathbf{Z}_1), \dots, \mathcal{M}(\mathbf{Z}_N)$ is bigger than or equal to $\min\{x, \hat{\gamma}_{t-1} + \delta\}$:
 - (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 (2.17), with \mathbf{v}_1 replaced by $\hat{\mathbf{v}}_T$.

By Proposition 5.2, at some iteration T we have $\hat{\gamma}_T \geq x$ and thus in step 5 we set $\hat{\gamma}_T := x$. It follows that we can view $\hat{\mathbf{v}}_T$ as the solution of a sample-average approximation problem, where the “true” problem is the CE maximization problem in (5.2). This is summarized in the following proposition.

Proposition 5.3 *Suppose that Assumption A holds. Then, Algorithm 5.2 converges w.p.1 to a solution of (2.21) after a finite number of iterations.*

We can then compare the approximating solution $\hat{\mathbf{v}}_T$ and the “true” solution \mathbf{v}^* using the asymptotic analysis for optimal solutions of stochastic optimization problems discussed in [26].

Following that approach, we obtain initially a *consistency* result: as $N \rightarrow \infty$, the distance between $\hat{\mathbf{v}}_T$ and the solution set defined in (5.2) goes to zero (w.p.1) provided that: i) the function $\ln f(\mathbf{z}, \mathbf{v})$ is continuous in \mathbf{v} , ii) the set V defined in assumption A is compact, and iii) there exists a function $h(\mathbf{z})$ such that $\mathbb{E}_{\mathbf{u}}[h(\mathbf{Z})] < \infty$ and $|\ln f(\mathbf{z}, \mathbf{v})| \leq h(\mathbf{z})$ for all \mathbf{z} and all $\mathbf{v} \in U$. We can also obtain a *distributional* result: under some additional assumptions involving the first- and second-order derivatives of $\ln f(\mathbf{z}, \mathbf{v})$ with respect to \mathbf{v} , and assuming uniqueness of \mathbf{v}^* in (5.2), it is possible to show that $N^{1/2}(\hat{\mathbf{v}}_T - \mathbf{v}^*) \xrightarrow{d} \mathcal{N}(0, B^{-1}\Sigma B^{-1})$, where \xrightarrow{d} indicates convergence in distribution, \mathcal{N} denotes the normal distribution, B is the Hessian matrix $\nabla_{\mathbf{v}}^2 D(\mathbf{u}, \mathbf{v}^*)$, and Σ is the covariance matrix given by

$$\mathbb{E}_{\mathbf{u}} \left[I_{\{\mathcal{M}(\mathbf{Y}) \geq x\}} \nabla_{\mathbf{v}} \ln f(\mathbf{Y}, \mathbf{v}^*) \nabla_{\mathbf{v}} \ln f(\mathbf{Y}, \mathbf{v}^*)' \right].$$

Notice that, in principle, this allows for the computation of $(1 - \alpha)\%$ -confidence regions for \mathbf{v}^* . We refer to [26] for details. Our experience, however, is that in the current rare-event setting such confidence regions are often too large for practical use. Notice also that one can use standard results from quantile estimation to derive $(1 - \alpha)\%$ -confidence intervals for $\gamma(\mathbf{v}, \rho)$; see, for instance, [3].

As discussed earlier, instead of using a constant δ in Algorithm 5.2 one can also apply a dynamic procedure where δ is changed according to the differences in $\hat{\gamma}_t$. More specifically, we can set $\delta_t := \hat{\gamma}_{t-1} - \hat{\gamma}_{t-2}$ between steps 5 and 6 of the algorithm, and use δ_t in step 6. The idea behind that is to update ρ_t as soon as the differences in the $\hat{\gamma}$'s start decreasing. Such procedure does not affect theoretical convergence and typically makes $\hat{\gamma}_t$ reach x faster; however, in some instances this might cause $\hat{\gamma}_t$ to increase "too fast", which in turn will yield a poor estimate $\hat{\mathbf{v}}_T$ in step 2 of the algorithm, due to the (relative) rarity of the event $\{\mathcal{M}(\mathbf{Z}) > \hat{\gamma}_{t-1}\}$ in (3.14). Thus, the more conservative choice $\delta = \text{constant}$ is recommended unless some pilot studies can be performed.

An even more conservative approach is to take $\delta = 0$ until the sequence $\{\hat{\gamma}_t\}$ gets "stalled", at which point a positive δ is used again. This approach yields the slowest progression of $\hat{\gamma}_t$, but in turn the estimate $\hat{\mathbf{v}}_T$ is more reliable. Notice however that, even if the optimal \mathbf{v}^* could be obtained, some problems might still require a very large sample size in (2.17); see the discussion in section 4. Given the limitations of one's computational budget, Algorithm 5.2 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).

5.1 The Root-Finding Problem

We briefly discuss now an application of the CE method to *root finding*. In many practical situations we need to estimate, for given ℓ , the root x^* of the non-linear equation

$$P_{\mathbf{u}}(\mathcal{M}(\mathbf{Y}) \geq x^*) = \mathbb{E}_{\mathbf{u}} \left[I_{\{\mathcal{M}(\mathbf{Y}) \geq x^*\}} \right] = \ell \quad (5.4)$$

rather than estimate ℓ itself. The estimate of x^* in (5.4) is based on the sample equivalent of $\mathbb{E}_{\mathbf{u}}[I_{\{\mathcal{M}(\mathbf{Y}) \geq x^*\}}] = \ell$ and can be obtained, for example, via stochastic approximation [25].

Alternatively, one can obtain x^* using the CE method. For example, Algorithm 5.2 can be modified as follows: first, remove step 5 of the algorithm and replace the min operation in step 6 with $\hat{\gamma}_{t-1} + \delta$; now, of course, $\hat{\gamma}_t$ grows unbounded. As soon as the probability $P_{\mathbf{u}}(\mathcal{M}(\mathbf{Y}) \geq \hat{\gamma}_t)$ — which is estimated by $\frac{1}{N} \sum_{i=1}^N I_{\{\mathcal{M}(\mathbf{Z}_i) \geq \hat{\gamma}_t\}} W(\mathbf{Z}_i, \mathbf{u}, \hat{\mathbf{v}}_t)$, where $\mathbf{Z}_1, \dots, \mathbf{Z}_N$ is a sample from $f(\mathbf{z}, \hat{\mathbf{v}}_t)$ — becomes smaller than ℓ , we know we have over-estimated x . We can then narrow down the search by a fine-tuning of $\hat{\gamma}_t$, using successively small values of δ in step 6 of the algorithm. Preliminary numerical experimentation suggests that such procedure works well.

The above root-finding procedure can also be used in connection with combinatorial optimization problems. Suppose that the underlying function $\mathcal{M}(\mathbf{Y})$ has finite support, and let x^* denote its (unknown) maximum value over the support. Suppose that assumption A holds with x^* in place of x . Then, using the root-finding procedure — i.e., Algorithm 5.2 without step 5 and without the min operation in step 6 — we have that $\hat{\gamma}_t$ grows until it surpasses x^* (which is easy to detect since in that case $P_{\mathbf{u}}(\mathcal{M}(\mathbf{Y}) \geq \hat{\gamma}_t) = 0$). Proposition 5.2 guarantees that such event occurs after a *finite* number of iterations. One can then perform the aforementioned fine-tuning procedure to determine x^* . We refer to [23] for details.

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 5.2 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 5.2). In these examples, such δ sufficed, i.e., the sequence $\{\hat{\gamma}_t\}$ never got stalled. Moreover, step 6(c) of Algorithm 5.2 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 , ($j = 1, \dots, J$) and a station k , ($k = 1, \dots, K$), 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 $\mathbf{Y} := (Y_{11}, \dots, Y_{KJ})$ 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 \mathbf{Y} , and hence is random. The above model is studied in [15] 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 $\mathcal{M}(\mathbf{Y}) = C_{KJ}(\mathbf{Y})$, we want to estimate $\ell(x) = P(\mathcal{M}(\mathbf{Y}) \geq x)$. Calculation of $\mathcal{M}(\mathbf{Y})$ for a particular realization of \mathbf{Y} can be done via the recursive formula

$$C_{kj} = \max(C_{k-1,j}, C_{k,j-1}) + Y_{kj}, \quad j = 1, \dots, J, \quad k = 1, \dots, K, \quad (6.1)$$

with $C_{k0} = C_{0j} = 0$, $k = 1, \dots, K$, $j = 1, \dots, J$. 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 [15] 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 independent, and that the service times of all jobs at a given machine have the same distribution.

We adopted the following methodology. First, for fixed J and K , we generated a problem randomly. This was accomplished by generating K mean service times μ_1, \dots, μ_K (one for each machine) uniformly between 10 and 40. We then estimated $P(\mathcal{M}(\mathbf{Y}) \geq x)$ for three values of x , based on the value of the total mean service time $\Gamma = J \sum_{k=1}^K \mu_k$. We took $x = 0.8\Gamma$, $x = \Gamma$ and $x = 2\Gamma$. 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 $P(\mathcal{M}(\mathbf{Y}) \geq \Gamma)$ should be small.

To estimate $P(\mathcal{M}(\mathbf{Y}) \geq x)$, 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 below displays the results for $J = 10$ jobs and $K = 5$ machines, which corresponds to 50 random variables. The mean service times for this data set are respectively 40, 36, 24, 38 and 14 for each machine. Although these results correspond to a particular instance of data, we must emphasize that the same type of results were observed for other problems we generated (for the same J and K and the same rule that mean service times are uniform between 10 and 40). Therefore, we report only one representative of the group. In the table, $\ell_N(x)$ is the estimate for $P(\mathcal{M}(\mathbf{Y}) \geq x)$, “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 each row is 96.67%; by Bonferroni’s inequality, the *overall* confidence on those intervals is at least 90%.

x	MC			CE		
	$\ell_N(x)$	90% H.W.	N	$\ell_N(x)$	90% H.W.	N
1216	8.973×10^{-4}	9.761×10^{-5}	4380 (± 10)	6.768×10^{-4}	9.463×10^{-5}	1000
1520	6.379×10^{-6}	5.202×10^{-6}	10983 (± 17)	3.329×10^{-6}	4.698×10^{-7}	2000
3040	0.000	0.000	40104 (± 683)	6.180×10^{-20}	5.091×10^{-20}	5000

Table 1: Estimated probabilities for exponential distribution case, $J = 10$, $K = 5$, $\mu_k \sim U(10, 40)$.

In order to do some comparisons with analytical calculations, we also studied the case where all service times have the same exponential distribution with mean μ . It is easy to check from (6.1)

that

$$C_{KJ} \geq Y_{11} + \dots + Y_{1J} + Y_{2J} + \dots + Y_{KJ} \sim \text{Gamma}(K + J - 1, \mu),$$

so that $P(\text{Gamma}(K + J - 1, \mu) \geq x)$ provides a lower bound on $P(\mathcal{M}(\mathbf{Y}) \geq x)$. To obtain an upper bound, we consider the Chebyshev inequality

$$P(\mathcal{M}(\mathbf{Y}) \geq x) \leq \frac{\mathbb{E}_\mu [\mathcal{M}(\mathbf{Y})^p]}{x^p}, \quad (6.2)$$

which is valid for any $p > 0$ (note that $\mathcal{M}(\mathbf{Y}) \geq 0$ in this example).

Table 2 below displays the estimation results for $J = 10$, $K = 5$, $\mu = 25$, whereas Table 3 lists the lower and upper bounds.

x	MC			CE		
	$\ell_N(x)$	90% H.W.	N	$\ell_N(x)$	90% H.W.	N
1000	7.715×10^{-5}	2.719×10^{-5}	4751 (± 117)	4.964×10^{-5}	1.044×10^{-5}	1000
1250	0.000	0.000	12931 (± 176)	2.679×10^{-8}	9.687×10^{-9}	2000
2500	0.000	0.000	46430 (± 786)	2.987×10^{-27}	2.903×10^{-27}	5000

Table 2: Estimated probabilities for exponential distribution case, $J = 10$, $K = 5$, $\mu = 25$.

x	lower bound	upper bound (95% H.W.)	p
1000	6.675×10^{-7}	1.148×10^{-3} (4.113×10^{-4})	35
1250	5.065×10^{-10}	6.853×10^{-7} (5.294×10^{-7})	35
2500	6.855×10^{-28}	5.263×10^{-25} (4.651×10^{-25})	60

Table 3: Estimated bounds for exponential distribution case, $J = 10$, $K = 5$, $\mu = 25$.

Finally, we studied the case where $K = 1$. In this case, the completion time is simply a sum of J i.i.d. exponentials with mean μ and therefore has Gamma distribution, so the probabilities can be computed analytically. In this case we took $x = 2\Gamma$, $x = 4\Gamma$ and $x = 6\Gamma$, where $\Gamma = J\mu$. Table 4 below displays the results for $J = 10$, $\mu = 25$. The true values for $x = 2\Gamma$, $x = 4\Gamma$ and $x = 6\Gamma$ are respectively 5×10^{-3} , 3.926×10^{-9} , and 3.331×10^{-16} , so we can see that the estimates obtained with the CE method are fairly close to the real values.

To illustrate the behavior of the algorithm, we also considered a smaller problem with $J = 5$ jobs, $K = 3$ machines, and all service times having the same exponential distribution with mean $\mu = 25$. The value of x chosen was $x = 1000$, for which the algorithm yielded the estimate

x	MC			CE		
	$\ell_N(x)$	90% H.W.	N	$\ell_N(x)$	90% H.W.	N
500	4.881×10^{-3}	2.650×10^{-4}	3220 (± 10)	4.917×10^{-3}	9.080×10^{-5}	1000
1000	0.000	0.000	10889 (± 20)	3.988×10^{-9}	7.352×10^{-11}	2000
1500	0.000	0.000	32477 (± 188)	2.906×10^{-16}	4.291×10^{-18}	5000

Table 4: Estimated probabilities for exponential distribution case, $J = 10$, $K = 1$, $\mu = 25$.

probability 1.590×10^{-10} ($\pm 4.076 \times 10^{-11}$) with 20 replications of sample size 1000 each. Table 5 displays, for each iteration t , the value of $\min\{x, \hat{\gamma}_{t-1}\}$ (computed in step 4 of Algorithm 5.2), as well as the new mean of each service time Y_{kj} , as calculated in step 2 of the algorithm. Notice that $\hat{\gamma}_t$ reaches x after 4 iterations.

t	$\min\{x, \hat{\gamma}_{t-1}\}$	$\hat{\mathbf{v}}_t$				
1	333.51	43.96	42.89	34.39	32.43	32.29
		34.59	36.50	42.75	37.52	34.03
		30.43	32.10	38.08	39.58	44.49
2	509.51	60.59	44.35	52.40	23.02	29.28
		69.70	54.73	56.20	79.37	38.69
		26.74	29.02	40.33	59.30	53.84
3	712.46	82.00	58.87	109.86	21.15	34.14
		59.76	83.10	61.76	102.86	34.23
		22.15	22.14	50.34	66.72	125.15
4	1000.0	181.73	46.79	128.56	26.48	23.45
		53.65	62.43	103.84	164.19	29.99
		13.01	26.12	94.70	73.22	191.68

Table 5: Progression of the algorithm for the exponential distribution case, $J = 5$, $K = 3$, $\mu = 25$.

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 independent, and that the service times of all jobs at a given machine have the same distribution.

Our data set was set up in the following way. For fixed J , K and m we generated, for each of the K 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 (cf. [15]). However, such procedure does not determine the probability of the maximum value, unless there is a single path corresponding to it. We then estimated $P(\mathcal{M}(\mathbf{Y}) \geq x)$ 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, $P(\mathcal{M}(\mathbf{Y}) > \Psi) = 0$).

To estimate $P(\mathcal{M}(\mathbf{Y}) \geq x)$, 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 7 below displays the results for $J = 10$ jobs, $K = 5$ machines and $m = 4$ possible outcomes for each random variable. This corresponds to 50 random variables and a 200-dimensional parameter vector. The values $y_{k\ell}$ taken on by each service time and the respective probabilities $v_{k\ell}$ are listed on table 6 below. Notice that, in this particular case, the exact probability for $x = \Psi = 541$ 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 on table 7. The notation used is the same as for the previous tables. We can see that the estimate obtained with the CE method is fairly close to the real value.

k	y_{k1}	v_{k1}	y_{k2}	v_{k2}	y_{k3}	v_{k3}	y_{k4}	v_{k4}
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: Values taken on by the service times and corresponding probabilities, for the data set with discrete distributions.

For the sake of comparison, we also considered the case where all service time distributions are

	MC			CE		
x	$\ell_N(x)$	90% H.W.	N	$\ell_N(x)$	90% H.W.	N
486	2.295×10^{-2}	1.840×10^{-3}	635 (± 1)	2.218×10^{-2}	2.127×10^{-3}	100
541	0.000	0.000	10193 (± 7)	4.952×10^{-7}	3.257×10^{-7}	700

Table 7: Estimated probabilities for discrete distribution case, $J = 10$, $K = 5$, $m = 4$, random data set.

uniform on the set $\{10, 20, 30, 40\}$, with $J = 10$ and $K = 5$. It is easy to see that the total completion time corresponds to a sum of $K + J - 1$ service times, and since those are assumed to have the same distribution, it follows that the maximum possible completion time is $\Psi = (K + J - 1) \times 40 = 560$. Table 8 below displays the results. Notice that it is difficult to compute the exact probability in this case, since there are many paths corresponding to the maximum completion time Ψ . A lower bound for the probability is $(1/4)^{K+J-1} = 3.725 \times 10^{-9}$, while the upper bound computed from (6.2) with $p = 60$ is $8.430 \times 10^{-5} \pm 9.931 \times 10^{-6}$.

	MC			CE		
x	$\ell_N(x)$	90% H.W.	N	$\ell_N(x)$	90% H.W.	N
500	1.423×10^{-2}	1.421×10^{-3}	680 (± 40)	1.424×10^{-2}	2.621×10^{-3}	100
560	0.000	0.000	10192 (± 19)	7.004×10^{-7}	4.637×10^{-7}	700

Table 8: Estimated probabilities for discrete distribution case, $J = 10$, $K = 5$, $m = 4$, uniform distribution.

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.

We conclude this section illustrating the behavior of the algorithm for the discrete distribution case. Again for the sake of brevity, we consider a smaller problem with $J = 4$ jobs and $K = 3$ machines; the distribution of the service times is the same as in the first three rows of table 6. The maximum value achieved by $\mathcal{M}(\mathbf{Y})$ in this case is $x = 237$, for which the algorithm yielded the estimate probability $0.0035 (\pm 0.0002)$ with 20 replications of sample size 200 each (the exact value can be calculated as 0.0036). Table 9 displays, for each iteration t , the value of $\min\{x, \hat{\gamma}_{t-1}\}$ (computed in step 4 of Algorithm 5.2) and the updated probability of each value taken on by Y_{kj} , as calculated in step 2 of the algorithm. Notice that $\hat{\gamma}_t$ reaches x after 3 iterations. Notice

also the “degenerate” effect referred to in Proposition 4.1 — the (k, 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 [23] for details.

7 Concluding Remarks

We have presented an efficient cross-entropy method for estimation of rare-event probabilities in static models via simulation. Besides establishing some theoretical foundations for the proposed algorithm, we presented supportive numerical results suggesting that the method indeed accomplishes its goal. We hope this work will stimulate the use of such techniques by others in problems where the underlying distributions belong to some parametric family.

Some issues for further research remain. For example, it would be important to find conditions under which the asymptotic solutions (with respect to x) of the variance minimization and cross-entropy problems ((2.18) and (2.20), respectively) coincide. Another topic to be explored is the use of the CE approach in the nonparametric setting, in which case one needs to solve a *functional* problem in order to determine optimal distributions.

References

- [1] Ali, S.M. and S. D. Silvey, (1996). “A general class of coefficients of divergence of one distribution from another”, *Journal of the Royal Statistical Society, Series B* **28**, 131–142.
- [2] Asmussen, S. and R.Y. Rubinstein, (1995). “Complexity properties of steady-state rare-events simulation in queueing models”, *Advances in Queueing: Theory, Methods and Open Problems* (J. Dshalalow, editor), CRC Press, 429–462.
- [3] Conover, W.J., (1999). *Practical Nonparametric Statistics*, 3rd. ed., John Wiley & Sons, New York.
- [4] de Boer, P.T., (2000). *Analysis and efficient simulation of queueing models of telecommunications systems*, Ph.D. thesis, University of Twente.
- [5] 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, Haifa, Israel.

t	$\min\{x, \hat{\gamma}_{t-1}\}$	(k, j)	$\hat{\mathbf{v}}_t$			
1	216	(1,1)	0.000	0.000	0.218	0.782
		(2,1)	0.019	0.275	0.127	0.579
		(3,1)	0.131	0.338	0.042	0.489
		(1,2)	0.210	0.062	0.220	0.508
		(2,2)	0.016	0.272	0.107	0.605
		(3,2)	0.110	0.283	0.035	0.572
		(1,3)	0.262	0.077	0.241	0.420
		(2,3)	0.016	0.294	0.139	0.551
		(3,3)	0.079	0.202	0.032	0.687
		(1,4)	0.300	0.088	0.262	0.350
		(2,4)	0.024	0.318	0.132	0.527
		(3,4)	0.000	0.032	0.030	0.939
2	236	...				
3	237	(1,1)	0.000	0.000	0.000	1.000
		(2,1)	0.000	0.000	0.000	1.000
		(3,1)	0.131	0.338	0.042	0.489
		(1,2)	0.132	0.039	0.139	0.690
		(2,2)	0.000	0.000	0.000	1.000
		(3,2)	0.110	0.283	0.035	0.572
		(1,3)	0.262	0.077	0.241	0.420
		(2,3)	0.000	0.000	0.000	1.000
		(3,3)	0.079	0.202	0.032	0.687
		(1,4)	0.300	0.088	0.262	0.350
		(2,4)	0.000	0.000	0.000	1.000
		(3,4)	0.000	0.000	0.000	1.000

Table 9: Progression of the algorithm for the discrete distribution case, $J = 4$, $K = 3$, data set as in table 6.

- [6] Dembo, A. and O. Zeitouni, (1998). *Large Deviations Techniques and Applications*, 2nd. ed., Springer-Verlag, New York.
- [7] Garvels, M.J.J. and D.P. Kroese, (1998). “A Comparison of RESTART Implementations”, *Proceedings of the 1998 Winter Simulation Conference*, 601–609, Washington, DC.
- [8] Glasserman, P., P. Heidelberger, P. Shahabuddin, and T. Zajic (1999), “Multilevel Splitting for Estimating Rare Event Probabilities”, *Operations Research*, Vol. 47, No. 4, 585–600.
- [9] Glasserman, P., P. Heidelberger, and P. Shahabuddin, (1999), “Asymptotically Optimal Importance Sampling and Stratification for Pricing Path-Dependent Options”, *Mathematical Finance*, 9: 117–152.
- [10] Glynn, P.W. and D.L. Iglehart, “Importance Sampling for Stochastic Simulations”, *Management Science*, Vol. 35, No. 11, 1367–1392.
- [11] Görg, C., (1999). “Simulating Rare Event Details of ATMDelay Time Distributions with RESTART/LRE”, *Proceedings of the RESIM Workshop*, 11–12 March, 1999, University of Twente, The Netherlands.
- [12] Jorgensen, B., (1997). *The Theory of Dispersion Models*, Chapman and Hall, London.
- [13] Heidelberger, P., (1995). “Fast simulation of rare events in queueing and reliability models”. *ACM Transaction of Modeling and Computer Simulation*, Vol. 5, No. 1, 43–85.
- [14] Haraszti, Z. and J. Townsend, (1999). “Rare Event Simulation of Delay in Packet Switching Networks Using DPR-Based Splitting”, *Proceedings of the RESIM Workshop*, 11–12 March, 1999, 185–190, University of Twente, The Netherlands.
- [15] Homem-de-Mello, T., A. Shapiro, and M.L. Spearman, (1999). “Finding optimal material release times using simulation based optimization”, *Management Science*, Vol. 45, no. 1, 86–102.
- [16] Kahn, H. and T.E. Harris, (1951). “Estimation of Particle Transmission by Random Sampling”, National Bureau of Standards Applied Mathematics Series.
- [17] Kaniowski, Y.M., A.J. King and R.J.-B. Wets, (1995). “Probabilistic bounds (via large deviations) for the solutions of stochastic programming problems”, *Annals of Operations Research*, 56, 189–208.

- [18] Kapur J.N. and H.K. Kesavan, (1992). *Entropy Optimization Principles with Applications*, Academic Press.
- [19] Kovalenko, I. (1995). “Approximations of queues via small parameter method”, *Advances in Queueing: Theory, Methods and Open Problems* (J. Dshalalow, editor), CRC Press, 481–509.
- [20] Krivan, V. and R.Y. Rubinstein, (1994). “Polynomial and exponential time algorithms for estimation of rare events in queueing models”, *Frontiers in Queueing: Models and Applications in Science and Engineering* (J. Dshalalow, editor), CRC Press.
- [21] MacLachlan, G. and T. Krishnan, (1997) *The EM Algorithm and Extensions*, John Wiley, New York.
- [22] Rubinstein, R.Y., (1997). “Optimization of computer simulation models with rare events”, *European Journal of Operations Research* **99**, 89–112.
- [23] Rubinstein, R.Y., (1999). “The cross- entropy method for combinatorial and continuous optimization”, *Methodology and Computing in Applied Probability* **2**, 127–190.
- [24] Rubinstein, R.Y., (2002). “Cross-entropy and rare-events for maximal cut and bipartition problems”. Manuscript, Technion, Haifa, Israel, (to be published in IEEE, Transactions on Simulation).
- [25] Rubinstein, R.Y. and B. Melamed (1998). *Modern Simulation and Modeling* John Wiley & Sons, New York.
- [26] Rubinstein, R.Y. and A. Shapiro. (1993). *Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization via the Score Function Method*, John Wiley & Sons, New York.
- [27] Shahabuddin, P. (1995). “Rare Event Simulation of Stochastic Systems,” *Proceedings of the 1995 Winter Simulation Conference, Washington, D.C.*, IEEE Press, 178-185.
- [28] Shapiro, A. and T. Homem-de-Mello, (2000). “On the rate of convergence of optimal solutions of Monte Carlo approximations of stochastic programs”, *SIAM Journal on Optimization*, **11**, 70-86.
- [29] Villén-Altamirano, M. and J. Villén-Altamirano (1999), “About the Efficiency of RESTART”, *Proceedings of the RESIM '99 Workshop*, 99–128.