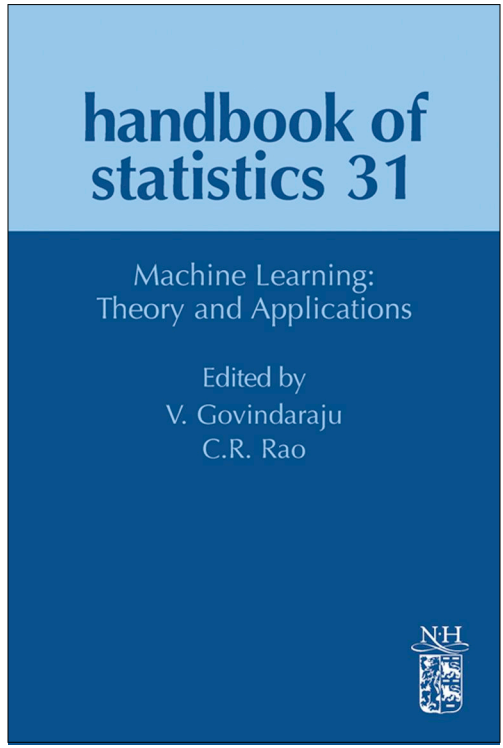

From Dr. Dirk Kroese, The Cross-Entropy Method for Estimation. In: Venu Govindaraju, C. R. Rao, editors, Handbook of Statistics, Vol 31. Chennai: Elsevier B.V., 2013, p. 19-34.
ISBN: 978-0-444-53859-8

Handbook of Statistics, *Vol. 31*
ISSN: 0169-7161

http://dx.doi.org/10.1016/B978-0-444-53859-8.00002-3

The Cross-Entropy Method for Estimation

Dirk P. Kroese[1], *Reuven Y. Rubinstein*[2], *and Peter W. Glynn*[3]

[1] *School of Mathematics and Physics, The University of Queensland, Brisbane 4072, Australia*
[2] *Faculty of Industrial Engineering and Management, Technion, Haifa, Israel*
[3] *Department of Management Science and Engineering, Stanford University, CA, USA*

Abstract

This chapter describes how difficult statistical estimation problems can often be solved efficiently by means of the *cross-entropy* (CE) method. The CE method can be viewed as an adaptive importance sampling procedure that uses the cross-entropy or Kullback–Leibler divergence as a measure of closeness between two sampling distributions. The CE method is particularly useful for the estimation of rare-event probabilities. The method can also be used to solve a diverse range of optimization problems. The optimization setting is described in detail in the chapter entitled "The Cross-Entropy Method for Optimization."

Keywords: cross-entropy, estimation, rare events, importance sampling, adaptive Monte Carlo, zero-variance distribution

1. Introduction

The CE method was introduced by Rubinstein (1999, 2001), extending earlier work on variance minimization (Rubinstein, 1997). Originally, the CE method was developed as a means of computing rare-event probabilities; that is, very small probabilities—say less than 10^{-4}. Naive Monte Carlo estimation of such a probability requires a large simulation effort, inversely proportional to the magnitude of the rare-event probability. The CE method is based on two ideas. The first idea is to estimate the probability of interest by gradually changing the sampling distribution, from the original to a distribution for which the rare event is much more likely to happen. To remove the estimation bias, importance sampling is used. The second idea is to use the CE distance to construct the sequence of sampling

distributions. This significantly simplifies the numerical computation at each step, and provides fast and efficient algorithms that are easy to implement by practitioners.

The CE method has been successfully applied to a diverse range of estimation and optimization problems. References on estimation include Asmussen et al. (2005), de Boer (2000), de Boer et al. (2004), Chan and Kroese (2010, 2012), Chan et al. (2011), Hui et al. (2005), Homem-de-Mello (2007), Kroese and Hui (2006), Kroese and Rubinstein (2004), Rao (2010), and Ridder (2005). Parallel implementations of the CE method are discussed in Evans (2009) and Evans et al. (2007), and other generalizations and advances are explored in Taimre (2009). For many more references on optimization we refer to the accompanying chapter in this handbook on CE optimization.

A tutorial on the CE method is given in de Boer et al. (2005). A comprehensive treatment can be found in Rubinstein and Kroese (2004); see also Rubinstein and Kroese (2007, Chapter 8) and Kroese et al. (2011, Chapter 13). The CE method homepage is www.cemethod.org.

2. Estimation setting

The general setting of the CE method concerns the estimation of an expectation ℓ of the form

$$\ell = \mathbb{E}_f[H(\mathbf{X})] = \int H(\mathbf{x}) f(\mathbf{x}) \mathrm{d}\mathbf{x}, \tag{1}$$

where H is a real-valued function and f is the probability density function (pdf) of a random variable (possibly multi-dimensional) $\mathbf{X}$. For simplicity it is assumed that $\mathbf{X}$ is a continuous random variable. For the discrete case, replace the integral in (1) with a sum. Of particular importance is the case where $H(\mathbf{x}) = \mathbf{I}_{\{S(\mathbf{x}) \geqslant \gamma\}}$, where S is another real-valued function (here and for the rest of the chapter $\mathbf{I}_A$ denotes an indicator function of an event A). This special case is further discussed in Section 2.4.

The *crude Monte Carlo* (CMC) estimator of ℓ is

$$\hat{\ell}_{\mathrm{CMC}} = \frac{1}{N} \sum_{k=1}^{N} H(\mathbf{X}_k),$$

where $\mathbf{X}_1, \ldots, \mathbf{X}_N$ are independent and identically distributed (iid) with density f. We write $\mathbf{X}_1, \ldots, \mathbf{X}_N \sim_{\mathrm{iid}} f$. It is not difficult to see (see, for example, Rubinstein and Kroese (2007, Page 27)) that the *relative error* of the CMC estimator is given by

$$\mathrm{RE}_{\mathrm{CMC}} \stackrel{\mathrm{def}}{=} \frac{\sqrt{\mathrm{Var}(\hat{\ell}_{\mathrm{CMC}})}}{\ell} = \frac{1-\ell}{N\ell} \approx \sqrt{\frac{1}{N\ell}}.$$

For example, to obtain a relative error of 1% for a probability of $\ell = 10^{-6}$, the sample size N needs to be 10^{10}.

The idea behind *importance sampling* is to sample the $\{\mathbf{X}_i\}$ from a different pdf g, under which the rare event may be more likely, while compensating for the bias

thus introduced. In particular, let g be a pdf for which $H(\mathbf{x})f(\mathbf{x}) \neq 0$ for every $\mathbf{x}$. The expectation ℓ can be written as

$$\ell = \int H(\mathbf{x}) \frac{f(\mathbf{x})}{g(\mathbf{x})} g(\mathbf{x}) \mathrm{d}\mathbf{x} = \mathbb{E}_g \left[H(\mathbf{X}) \frac{f(\mathbf{X})}{g(\mathbf{X})} \right], \tag{2}$$

where the subscript g indicates that the expectation is taken with respect to g rather than f. Consequently, if $\mathbf{X}_1, \ldots, \mathbf{X}_N \sim_{\text{iid}} g$, then

$$\hat{\ell} = \frac{1}{N} \sum_{k=1}^{N} H(\mathbf{X}_k) \frac{f(\mathbf{X}_k)}{g(\mathbf{X}_k)} \tag{3}$$

is an unbiased estimator of ℓ. The quotient $W(\mathbf{X}) \overset{\text{def}}{=} f(\mathbf{X})/g(\mathbf{X})$ is called the *likelihood ratio* of $\mathbf{X}$. Estimating ℓ via such an *importance sampling* estimator may be beneficial, in terms of a smaller relative error for the same simulation effort, if the importance sampling pdf g is chosen appropriately. Note that by taking $f = g$ one obtains the CMC estimator.

2.1. Variance minimization

The optimal importance sampling pdf, that is, the pdf g^* for which the variance of $\hat{\ell}$ is minimal, is proportional to $|H|f$ (see, e.g., Rubinstein and Kroese (2007, Page 132)). Indeed, if H is strictly positive then under $g^* \propto Hf$ the importance sampling estimator has *zero variance*. However, g^* is in general difficult to evaluate; for example, when $H \geqslant 0$ the normalization constant of g^* is precisely ℓ. Often the "nominal" pdf f can be embedded in a parameterized class of densities $\{f(\cdot; \mathbf{v})\}$, where f corresponds to some $f(\cdot; \mathbf{u})$. A sensible approach is to search for the best importance sampling pdf $f(\cdot; \mathbf{v})$; that is, find the parameter $\mathbf{v}$ for which the importance sampling estimator $\hat{\ell}$ has the smallest variance. Since

$$\begin{aligned} \mathrm{Var}_{\mathbf{v}}(\hat{\ell}) &= \frac{1}{N} \mathrm{Var}_{\mathbf{v}}(H(\mathbf{X}) W(\mathbf{X}; \mathbf{u}, \mathbf{v})) \\ &= \frac{1}{N} \left(\mathbb{E}_{\mathbf{v}} \left[H(\mathbf{X})^2 W(\mathbf{X}; \mathbf{u}, \mathbf{v})^2 \right] - \ell^2 \right), \end{aligned}$$

where $W(\mathbf{X}; \mathbf{u}, \mathbf{v}) = f(\mathbf{X}; \mathbf{u})/f(\mathbf{X}; \mathbf{v})$, the optimal $\mathbf{v}$ is found by solving the following *variance minimization program*:

$$\min_{\mathbf{v}} \mathbb{E}_{\mathbf{v}} \left[H(\mathbf{X})^2 W(\mathbf{X}; \mathbf{u}, \mathbf{v})^2 \right].$$

This is in general not an easy problem.

2.2. Cross-entropy minimization

The idea of the CE method is to choose the importance sampling pdf g in a specified class of pdfs such that the Kullback–Leibler divergence between the optimal importance sampling pdf g^* and g is minimal. The Kullback–Leibler divergence between two pdfs g and h is given by

$$\mathcal{D}(g,h) = \mathbb{E}_g\left[\ln \frac{g(\mathbf{X})}{h(\mathbf{X})}\right] = \int g(\mathbf{x}) \ln \frac{g(\mathbf{x})}{h(\mathbf{x})} \mathrm{d}\mathbf{x} \tag{4}$$
$$= \int g(\mathbf{x}) \ln g(\mathbf{x}) \mathrm{d}\mathbf{x} - \int g(\mathbf{x}) \ln h(\mathbf{x}) \mathrm{d}\mathbf{x}.$$

We assume from now on that the function H is positive, that the nominal pdf f is parameterized by a finite-dimensional vector $\mathbf{u}$; that is, $f(\mathbf{x}) = f(\mathbf{x};\mathbf{u})$, and that the importance sampling pdf is $f(\cdot;\mathbf{v})$ for some parameter $\mathbf{v}$. The CE minimization procedure involves finding an optimal reference parameter vector, $\mathbf{v}^*$ say, by cross-entropy minimization:

$$\begin{aligned}
\mathbf{v}^* &= \operatorname*{argmin}_{\mathbf{v}} \mathcal{D}(g^*, f(\cdot;\mathbf{v})) \\
&= \operatorname*{argmax}_{\mathbf{v}} \int H(\mathbf{x}) f(\mathbf{x};\mathbf{u}) \ln f(\mathbf{x};\mathbf{v}) \mathrm{d}\mathbf{x} \\
&= \operatorname*{argmax}_{\mathbf{v}} \mathbb{E}_{\mathbf{u}} H(\mathbf{x}) \ln f(\mathbf{X};\mathbf{v}) \\
&= \operatorname*{argmax}_{\mathbf{v}} \mathbb{E}_{\mathbf{w}} H(\mathbf{x}) W(\mathbf{X};\mathbf{u},\mathbf{w}) \ln f(\mathbf{X};\mathbf{v}),
\end{aligned} \tag{5}$$

where $\mathbf{w}$ is any reference parameter. This $\mathbf{v}^*$ can be estimated via the stochastic counterpart of (5):

$$\hat{\mathbf{v}} = \operatorname*{argmax}_{\mathbf{v}} \frac{1}{N} \sum_{k=1}^{N} H(\mathbf{X}_k) W(\mathbf{X}_k;\mathbf{u},\mathbf{w}) \ln f(\mathbf{X}_k;\mathbf{v}), \tag{6}$$

where $\mathbf{X}_1,\ldots,\mathbf{X}_N \sim_{\text{iid}} f(\cdot;\mathbf{w})$. This leads to the following algorithm.

Algorithm 2.1 (CE Algorithm)

1. Choose a reference parameter $\mathbf{w}$, *e.g.*, $\mathbf{w} = \mathbf{u}$, and generate $\mathbf{X}_1,\ldots,\mathbf{X}_N \sim_{\text{iid}} f(\cdot;\mathbf{w})$.
2. Estimate the CE optimal parameter $\hat{\mathbf{v}}$ from the stochastic program (6).
3. Generate $\mathbf{X}_1,\ldots,\mathbf{X}_{N_1} \sim_{\text{iid}} f(\cdot;\hat{\mathbf{v}})$ and estimate ℓ via importance sampling, as in (3).

It can be beneficial to iterate Steps 1 and 2 a number of times in Algorithm 2.1; that is, in subsequent iterations the parameter $\mathbf{w}$ in Step 1 is chosen as $\hat{\mathbf{v}}$, obtained in Step 2.

2.3. *Updating formulae*

The main benefit of using CE program (6) over the corresponding variance minimization program is that the former yields simple explicit solutions for certain important cases. The optimization of (6) is similar to the calculation of the maximum likelihood estimator of $\mathbf{v}$ for the pdf $f(\cdot;\mathbf{v})$. Indeed, by taking the gradient of the sum in (6) and equating it to the zero vector $\mathbf{0}$, one obtains $\hat{\mathbf{v}}$ (under mild regularity conditions) as the solution to the equation

$$\frac{1}{N} \sum_{k=1}^{N} H(\mathbf{X}_k) W(\mathbf{X}_k;\mathbf{u},\mathbf{w}) \nabla \ln f(\mathbf{X}_k;\mathbf{v}) = \mathbf{0}, \tag{7}$$

where $\nabla \ln f(\mathbf{X}; \mathbf{v})$ is the well-known *score function* in maximum likelihood estimation. We discuss two important special cases.

2.3.1. Exponential families

Suppose that $\{f(\cdot; \eta)\}$ forms an m-dimensional exponential family in natural parameter space; that is,

$$f(\mathbf{x}; \boldsymbol{\eta}) = c(\boldsymbol{\eta}) \mathrm{e}^{\boldsymbol{\eta}^\top \mathbf{t}(\mathbf{x})} h(\mathbf{x}), \tag{8}$$

where $\mathbf{t}(\mathbf{x}) = (t_1(\mathbf{x}), \ldots, t_m(\mathbf{x}))^\top, c(\boldsymbol{\eta}) > 0$, and $h(\mathbf{x}) > 0$, for all $\boldsymbol{\eta} = (v_1, \ldots, v_m)^\top$ such that

$$A(\boldsymbol{\eta}) \stackrel{\text{def}}{=} -\ln c(\boldsymbol{\eta}) = \ln \int \mathrm{e}^{\boldsymbol{\eta}^\top \mathbf{t}(\mathbf{x})} h(\mathbf{x}) \mathrm{d}\mathbf{x} < \infty.$$

(Replace the integral by a sum in the discrete case.)

The random vector $\mathbf{t}(\mathbf{X})$ is a sufficient statistic for $\boldsymbol{\eta}$. Moreover, $\mathbb{E}_{\boldsymbol{\eta}} \mathbf{t}(\mathbf{X}) = \nabla A(\boldsymbol{\eta})$ and $\mathrm{Cov}_{\boldsymbol{\eta}}(\mathbf{t}(\mathbf{X})) = \nabla^2 A(\boldsymbol{\eta})$. The score function becomes $\nabla \ln f(\mathbf{x}; \boldsymbol{\eta}) = \mathbf{t}(\mathbf{x}) - \nabla A(\boldsymbol{\eta})$. It follows that the solution $\hat{\boldsymbol{\eta}}$ to (7) (replacing $\mathbf{u}, \mathbf{v}, \mathbf{w}$ with $\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\omega}$) satisfies

$$\nabla A(\hat{\boldsymbol{\eta}}) = \frac{\sum_{k=1}^N H(\mathbf{X}_k) W(\mathbf{X}_k; \boldsymbol{\theta}, \boldsymbol{\omega}) \mathbf{t}(\mathbf{X}_k)}{\sum_{k=1}^N H(\mathbf{X}_k) W(\mathbf{X}_k; \boldsymbol{\theta}, \boldsymbol{\omega})}. \tag{9}$$

In particular, if the exponential family is reparameterized by the mean, $\mathbf{v} = \nabla A(\boldsymbol{\eta})$, then the CE optimal parameter $\hat{\mathbf{v}}$ is explicitly given in the right-hand side of (9).

Example 2.1 (Exponential Random Variables) Let $X_1, \ldots, X_m$ be independent and identically distributed random variables with parameter θ. We write $X_1, \ldots, X_m \sim_{\text{iid}} \mathsf{Exp}(\theta)$. Let η be the reference parameter of the importance sampling pdf $f(\mathbf{x}; \eta)$ given by

$$f(\mathbf{x}; \eta) = \prod_{i=1}^m \eta \, \mathrm{e}^{-x_i \eta} = \eta^m \, \mathrm{e}^{-\eta \sum_{i=1}^m x_i},$$

which is a one-dimensional exponential family of the form (8), with $\mathbf{t}(\mathbf{x}) = -\sum_{i=1}^m x_i$ and $c(\eta) = \eta^m$. Note that under this importance sampling pdf, $X_1, \ldots, X_n \sim_{\text{iid}} \mathsf{Exp}(\eta)$. Writing $H_k = H(\mathbf{X}_k)$ and the likelihood ratio $W_k = f(\mathbf{X}_k; \theta)/f(\mathbf{X}_k; \omega)$ in (6), the CE optimal parameter $\hat{\eta}$ is found from

$$-\frac{\mathrm{d} \ln c(\eta)}{\mathrm{d}\eta} = \frac{m}{\eta} = \frac{\sum_{k=1}^N H_k W_k \sum_{i=1}^m X_{ki}}{\sum_{k=1}^N H_k W_k},$$

where X_{ki} is the ith component of $\mathbf{X}_k$. Reparameterizing the $\mathsf{Exp}(\eta)$ distribution by the mean $v = 1/\eta$, the CE updating formula for v thus becomes:

$$\hat{v} = \frac{\sum_{k=1}^N H_k W_k \sum_{i=1}^m X_{ki}/m}{\sum_{k=1}^N H_k W_k}, \quad i = 1, \ldots, n. \tag{10}$$

2.3.2. Discrete distributions

Suppose $X_1, \ldots, X_n$ are iid random variables taking values $1, \ldots, m$ with probabilities $v_1, \ldots, v_m$, respectively. Let $\mathbf{v} = (v_1, \ldots, v_{m-1})$. Note that $v_m = (1 - v_1 - \cdots - v_{m-1})$. The (discrete) pdf of each component X is given by

$$f(x; \mathbf{v}) = \prod_{i=1}^{m} v_i^{\mathrm{I}_{\{x=i\}}} = \left(1 - \sum_{i=1}^{m-1} v_i\right)^{1-\sum_{i=1}^{m-1} \mathrm{I}_{\{x=i\}}} \times \prod_{i=1}^{m-1} v_i^{\mathrm{I}_{\{x=i\}}},$$

The $m-1$ components of the corresponding score function are given by

$$\frac{\mathrm{d} \ln f(x; \mathbf{v})}{\mathrm{d} v_i} = \frac{\mathrm{I}_{\{x=i\}}}{v_i} - \frac{\mathrm{I}_{\{x=m\}}}{v_m}, \quad i = 1, \ldots, m-1.$$

Substitution in (7), and using the abbreviations $H_k = H(X_k)$ and $W_k W(X_k; \mathbf{u}, \mathbf{w})$, shows that

$$\hat{v}_i = \frac{\sum_{k=1}^{N} H_k W_k \mathrm{I}_{\{X_k=i\}}}{\sum_{k=1}^{N} H_k W_k \mathrm{I}_{\{X_k=m\}}} \hat{v}_m.$$

The parameter $\hat{v}_m$ can be found from the fact that $\hat{v}_1 + \cdots + \hat{v}_m = 1$. It follows that

$$\hat{v}_i = \frac{\sum_{k=1}^{N} H_k W_k \mathrm{I}_{\{X_k=i\}}}{\sum_{k=1}^{N} H_k W_k}, \quad i = 1, \ldots, m.$$

2.4. Rare-event simulation

Often the quantity of interest $\ell = \mathbb{E} H(\mathbf{X})$ is a probability of the form $\ell = \mathbb{P}(S(\mathbf{X}) \geqslant \gamma)$ for some function S and *level* γ; that is, $H(\mathbf{x}) = \mathrm{I}_{\{S(\mathbf{X}) \geqslant \gamma\}}$. A complication in solving (6) or (7) occurs when ℓ is a rare-event probability. In that case the optimization program and the updating formula becomes useless for moderate sample size N, because all (or most) of the values $H(\mathbf{X}_k)$ are zero. One remedy is to use a *multi-level* CE procedure instead, where a sequence of reference parameters and levels is constructed with the goal that the former converges to $\mathbf{v}^*$ and the latter to γ. The idea is to first choose a level parameter $\hat{\ell}_1$ for which the event $S(\mathbf{X}) \geqslant \hat{\gamma}_1$ is not too rare under the original pdf $f(\cdot; \mathbf{u})$, and then estimate the CE optimal parameter, say $\hat{v}_1$ for this level via (6) (which is now not devoid of meaning). Specifically, $\hat{\gamma}_1$ is chosen as the sample $(1-\varrho)$-quantile of $S(\mathbf{X})$, based on a random sample $\mathbf{X}_1, \ldots, \mathbf{X}_N \sim_{\mathrm{iid}} f(\cdot; \mathbf{u})$. This procedure is then iterated by choosing $\hat{\gamma}_2$ as the sample $(1-\varrho)$-quantile of $S(\mathbf{X})$, based on a random sample from $f(\cdot; \hat{\mathbf{v}}_1)$, estimating the optimal CE parameter $\hat{\mathbf{v}}_2$ from (6), and so on. The complete procedure is summarized as follows; see, e.g., Rubinstein and Kroese (2007, Page 238).

Algorithm 2.2 (CE Algorithm for Rare-Event Estimation)

1. Define $\hat{\mathbf{v}}_0 = \mathbf{u}$. Let $N^{\mathrm{e}} = \lceil \varrho N \rceil$. Set $t = 1$ (iteration counter).
2. Generate $\mathbf{X}_1, \ldots, \mathbf{X}_N \sim_{\mathrm{iid}} f(\cdot; \hat{\mathbf{v}}_{t-1})$. Calculate $S_i = S(\mathbf{X}_i)$ for all i, and order these from smallest to largest: $S_{(1)} \leqslant \cdots \leqslant S_{(N)}$. Let $\hat{\gamma}_t$ be the sample $(1-\varrho)$-quantile of performances; that is, $\hat{\gamma}_t = S_{(N-N^{\mathrm{e}}+1)}$. If $\hat{\gamma}_t > \gamma$, reset $\hat{\gamma}_t$ to γ.

3. Use the **same** sample $\mathbf{X}_1, \ldots, \mathbf{X}_N$ to solve the stochastic program (6), with $\mathbf{w} = \hat{\mathbf{v}}_{t-1}$. Denote the solution by $\hat{\mathbf{v}}_t$.
4. If $\hat{\gamma}_t < \gamma$, set $t = t + 1$ and reiterate from Step 2; otherwise, proceed with Step 5.
5. Let T be the final iteration counter. Generate $\mathbf{X}_1, \ldots, \mathbf{X}_{N_1} \sim_{\text{iid}} f(\cdot; \hat{\mathbf{v}}_T)$ and estimate ℓ via importance sampling, as in (3).

Apart from specifying the family of sampling pdfs, the sample sizes N and N_1, and the rarity parameter ϱ (typically between 0.01 and 0.1), the algorithm is completely self-tuning. The sample size N for determining a good reference parameter can usually be chosen much smaller than the sample size N_1 for the final importance sampling estimation, say $N = 1000$ versus $N_1 = 100{,}000$. Under certain technical conditions the deterministic version of Algorithm 2.2 is guaranteed to terminate (reach level γ) provided that ϱ is chosen small enough; see Section 3.5 of Rubinstein and Kroese (2004).

Example 2.2 (Stochastic Activity Network) Figure 1 shows an example of a *stochastic activity network*. Such networks are frequently used in process management to schedule concurrent activities of some project from start to finish. Each arc in the network corresponds to an activity, and is weighted by the duration of that activity. The nodes in the network represent milestones. Any activity originating from a milestone can only be started once all activities leading to that milestone have been completed. The maximal project duration corresponds to the length of the longest path in the graph. Figure 1 shows a stochastic activity network with 10 activities. Suppose the durations of the activities are independent exponential random variables $X_1, \ldots, X_{10}$, each with means 1.

Let $S(\mathbf{X})$ denote length of the longest path in the graph; that is,

$$S(\mathbf{X}) = \max_i L_i,$$

where

$$\begin{aligned} L_1 &= X_1 + X_4 + X_9, \\ L_2 &= X_3 + X_6 + X_9, \\ L_3 &= X_3 + X_8, \\ L_4 &= X_3 + X_7 + X_{10}, \\ L_5 &= X_2 + X_5 + X_{10}. \end{aligned}$$

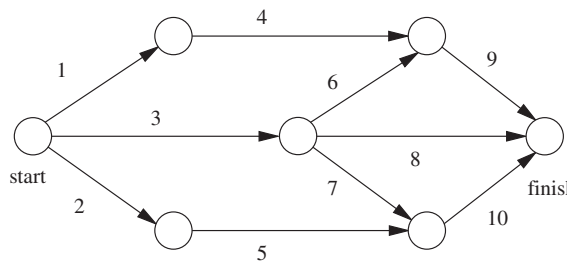

Fig. 1. A stochastic activity network.

Table 1
Convergence of the sequence $\{(\hat{\gamma}_t, \hat{\mathbf{v}}_t)\}$

t	$\hat{\gamma}_t$	$\hat{\mathbf{v}}_t$									
0	–	1	1	1	1	1	1	1	1	1	1
1	7.05	1.479	1.453	1.763	1.473	1.441	1.398	1.427	1.147	1.707	1.732
2	11.09	1.794	1.826	2.422	1.773	1.798	1.735	1.731	1.176	2.449	2.444
3	14.69	2.034	2.034	3.056	2.213	2.167	2.125	1.982	1.040	3.070	2.977
4	17.87	2.097	2.297	3.570	2.352	2.327	2.281	2.593	1.021	3.614	3.842
5	20.00	2.936	2.548	3.683	2.537	2.628	2.454	2.227	1.209	4.004	3.683

Suppose the objective is to estimate the rare-event probability $\mathbb{P}(S(\mathbf{X}) \geqslant 20)$ using importance sampling where the random vector $\mathbf{X} = (X_1, \ldots, X_{10})$ has independent exponentially distributed components with mean vector $\mathbf{v} = (v_1, \ldots, v_{10})$. Note that the nominal pdf is obtained by setting $v_i = 1$ for all i. At the tth iteration of the multi-level CE Algorithm 2.2, the solution to (6) with $H(\mathbf{X}) = \mathrm{I}_{\{S(\mathbf{X}) \geqslant \hat{\gamma}_t\}}$ is, similar to (10), given by

$$\hat{v}_{t,i} = \frac{\sum_{k=1}^{N} \mathrm{I}_{\{S(\mathbf{X}_k) \geqslant \hat{\gamma}_t\}} W_k X_{ki}}{\sum_{k=1}^{N} \mathrm{I}_{\{S(\mathbf{X}_k) \geqslant \hat{\gamma}_t\}} W_k}, \tag{11}$$

where $\mathbf{X}_1, \ldots, \mathbf{X}_N \sim_{\text{iid}} f(\cdot; \hat{\mathbf{v}}_{t-1})$, $W_k = f(\mathbf{X}_k; \mathbf{u})/f(\mathbf{X}_k; \hat{\mathbf{v}}_{t-1})$, and X_{ki} is the ith element of $\mathbf{X}_k$.

Table 1 lists the successive estimates for the optimal importance sampling parameters obtained from the multi-level CE algorithm, using $N = 10^5$ and $\varrho = 0.1$.

The last step in Algorithm 2.2 gives an estimate of $1.72 \cdot 10^{-6}$ with an estimated relative error of 2%, using a sample size of $N_1 = 10^6$. A typical crude Monte Carlo estimate (that is, taking $\mathbf{v} = \mathbf{u} = (1, 1, \ldots, 1)$) using the same sample size is $3 \cdot 10^{-6}$, with an estimated relative error of 60%, and is therefore of little use.

3. Extensions

3.1. Sampling directly from the zero-variance distribution

For certain rare-event estimation problems it is possible to (approximately) sample directly from the zero-variance importance sampling pdf g^*, for example via Markov chain Monte Carlo. By sampling directly from g^* one can estimate the CE optimal parameters in a single step and without likelihood ratios. The idea was first introduced in Chan (2010). This approach could be advantageous for high-dimensional problems, where the *likelihood degeneracy* can pose a serious impediment to importance sampling; see, for example, Rubinstein and Kroese (2007, Page 133).

To explain the idea, suppose $\ell = \mathbb{P}(S(\mathbf{X}) \geqslant \gamma)$ is the rare event of interest, where $\mathbf{X} \sim f(\cdot; \mathbf{u})$. The zero-variance importance sampling density g^* is simply the conditional pdf f given the event $S(\mathbf{X}) \geqslant \gamma$; that is,

$$g^*(\mathbf{x}) = \frac{f(\mathbf{x}; \mathbf{u}) \mathrm{I}_{\{S(\mathbf{x}) \geqslant \gamma\}}}{\ell}.$$

The CE optimal parameter $\mathbf{v}^*$ is

$$\begin{aligned}\mathbf{v}^* &= \underset{\mathbf{v}}{\operatorname{argmax}} \int \mathrm{I}_{\{S(\mathbf{x})\geqslant\gamma\}} f(\mathbf{x};\mathbf{u}) \ln f(\mathbf{x};\mathbf{v}) \mathrm{d}\mathbf{x} \\ &= \underset{\mathbf{v}}{\operatorname{argmax}} \mathbb{E}_{g^*} \ln f(\mathbf{x};\mathbf{v}) \mathrm{d}\mathbf{x},\end{aligned}$$

which can be estimated via the sample average approximation:

$$\hat{\mathbf{v}}^* = \underset{\mathbf{v}}{\operatorname{argmax}} \frac{1}{N} \sum_{k=1}^{N} \ln f(\mathbf{X}_k;\mathbf{v}), \tag{12}$$

where $\mathbf{X}_1,\ldots,\mathbf{X}_N$ is a (approximate) sample from g^*. This leads to the following algorithm; see Chan and Kroese (2011).

Algorithm 3.1 (CE Algorithm via the Zero-Variance Distribution)

1. Generate a random sample $\mathbf{X}_1,\ldots,\mathbf{X}_N$ from the density g^* and find the solution to (12).
2. Generate a sample $\mathbf{X}_1,\ldots,\mathbf{X}_M$ from the density $f(\cdot;\hat{\mathbf{v}}^*)$ and estimate ℓ via importance sampling, as in (3).

Note that no likelihood ratio or indicator is involved. As a result, the algorithm does not only afford substantial computational saving in high-dimensional settings, its solution is more robust and numerically stable as well. Generating draws from g^*, however, requires additional effort, but with the advent of Markov chain Monte Carlo (MCMC) methods, this problem is well studied and a variety of techniques are available to our disposal. In particular, sampling from the zero-variance pdf g^* can be achieved without knowledge of ℓ. The number of draws required to estimate $\hat{\mathbf{v}}^*$ is typically much smaller than that required in the multi-level CE algorithm. As (12) is a maximum likelihood type estimator, where sample is taken from g^* instead of $f(\cdot;\mathbf{v})$, the solution can often be obtained analytically. Note, finally, that in this approach the function S must be explicitly available, while in the standard CE method any importance sampling pdf can be used, in conjunction with the likelihood ratio.

Example 3.1 (Binomial Distribution) To compare the quality of the optimal reference parameter estimators for the multi-level CE Algorithm 2.2 and the zero-variance CE Algorithm 3.1, consider the estimation of $\mathbb{P}(S(\mathbf{X}) \geqslant \gamma)$, where $S(\mathbf{X}) = X_1 + \cdots + X_n$, $\gamma = n\beta$ for some $\beta \ln (0,1)$, and $X_i \sim \mathsf{Ber}(p_i)$, $i = 1,\ldots,n$, independently. The nominal density is thus $f(\mathbf{x};\mathbf{p}) = \prod_{i=1}^n p_i^{x_i}(1-p_i)^{1-x_i}$, where $\mathbf{x} = (x_1,\ldots,x_n)$ and $\mathbf{p} = (p_1,\ldots,p_n)$. We wish to locate the optimal importance density within the parametric family $f(\mathbf{x};\mathbf{q})$ indexed by $\mathbf{q} = (q_1,\ldots,q_n)$, where $q_i \in (0,1)$ for $i = 1,\ldots,n$. The CE optimal parameter $\mathbf{q}^*$ follows from the maximization program

$$\max_{\mathbf{q}} \sum_{\mathbf{x}:S_n(\mathbf{x})\geqslant\gamma} \left(\prod_{i=1}^n p_i^{x_i}(1-p_i)^{(1-x_i)}\right)\left(\sum_{i=1}^n x_i \ln q_i + (1-x_i)\ln(1-q_i)\right),$$

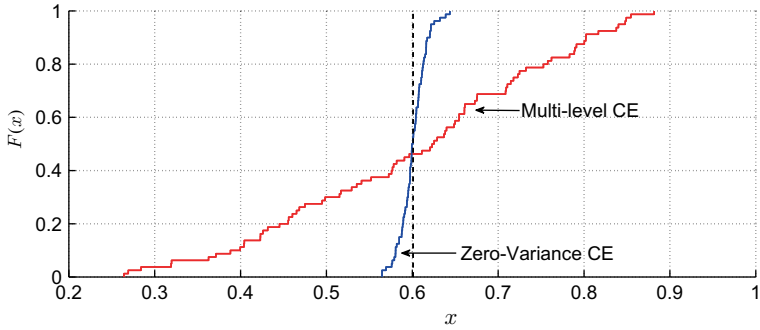

Fig. 2. The empirical distribution function of the CE estimates.

which yields the closed-form expression

$$q_j^* = \frac{\sum_{\mathbf{x}:S_n(\mathbf{x})\geqslant\gamma} x_j \prod_{i=1}^n p_i^{x_i}(1-p_i)^{1-x_i}}{\sum_{\mathbf{x}:S_n(\mathbf{x})\geqslant\gamma} \prod_{i=1}^n p_i^{x_i}(1-p_i)^{1-x_i}}, \quad j = 1,\ldots,n.$$

In particular, if all the p_i are identical, then $q_j^* = \lceil \beta n \rceil / n$. We estimate $\mathbf{q}^*$ via the multi-level CE procedure and by sampling from the zero-variance pdf. As a numerical example, we first set $n = 80$, $\beta = 0.6$, and $p_1 = \cdots = p_n = 0.1$. For the multi-level CE method, we set $N = 10{,}000$ and $\rho = 0.01$. The algorithm terminates at the fifth iteration, requiring a total of 50,000 draws. For the zero-variance CE procedure, we run a Gibbs sampler with 10 parallel chains, each has a length of 1000, and the total budget is therefore 10,000. It is also worth mentioning that drawing from g^* via the Gibbs sampler in this case only requires generating Bernoulli draws. The empirical cumulative distribution functions (cdf) of the CE and "zero-variance CE" estimates, together with the optimal reference parameter calculated analytically, are presented in Fig. 2.

It is evident that the multi-level CE estimates fluctuate more widely compared to those obtained by the zero-variance version, even though the simulation budget for the former is five times as large. Using a sample size of $N = 50{,}000$, typical importance sampling estimates for the multi-level and zero-variance approaches are $3 \cdot 10^{-28}$ and $7.96 \cdot 10^{-28}$, respectively, with estimated relative errors 0.3 and 0.01. The true probability is approximately $8.10 \cdot 10^{-28}$. It is important to note that the multi-level CE estimation procedure can be considerably improved by simply increasing the sample size for estimating the parameters, for example from 10,000 to 100,000. Similar experiments indicate that as the dimension n of the problem gets larger, the multi-level CE estimates become more unreliable, compared to the zero-variance counterpart, which are essentially unaffected by the increase in dimension.

The result from this toy example suggests a reason why the multi-level CE method fails to give accurate estimates in high-dimensional settings: the reference parameter vector obtained is suboptimal, and therefore the resulting importance density does not sufficiently mimic the behavior of g^*. In principle one can increase the accuracy of the multi-level CE estimates by increasing the sample size N or the rarity parameter ϱ. In either case, however, the total simulation effort would increase, and

in moderately high-dimensional problems, this approach might not be practical. On the other hand, the result also suggests that if we avoid the multi-level maximization procedure and estimate $\mathbf{v}^*$ directly via (12), we can improve the performance of the standard CE procedure.

Example 3.2 (Stochastic Activity Network (Continued)) This is to show how the zero-variance CE approach can be applied to the stochastic activity network of Example 2.2.

We can sample from the zero-variance pdf g^* using the Gibbs sampler. This requires sampling from the marginal distributions of g^* conditional on the values of the other components. Consider first sampling X_1 conditional on $x_2,\ldots,x_{10}$. If any of the lengths $L_i, i = 2,\ldots,5$ is greater than or equal to γ, then the marginal distribution of X_1 given $x_2,\ldots,x_{10}$ is simply equal to the nominal distribution (that is, $\mathsf{Exp}(1)$). However, if every $L_2,\ldots,L_{10}$ is less than γ, then in order for L_1 to exceed γ, X_1 has to be greater than or equal to $\min\{\gamma - x_4 - x_6, 0\}$. Hence, to sample from the marginal conditional distribution in this case, draw $Z \sim \mathsf{Exp}(1)$ and compute $X_1 = Z + \min\{\gamma - x_4 - x_6, 0\}$. That is X_1 is drawn from a truncated exponential distribution. Similarly, given $x_1, x_3, \ldots, x_{10}, X_2$ is drawn from its original distribution if $\max_{i\leqslant 4} L_i \geqslant \gamma$; otherwise, set $X_2 = \min\{\gamma - x_5 - x_{10}, 0\} + Z$, where $Z \sim \mathsf{Exp}(1)$. More precisely, the Gibbs sampling procedure for this problem is as follows.

Algorithm 3.2 (Gibbs Sampling for the Stochastic Activity Network)

1. Initialize $\mathbf{X}$ such that $S(\mathbf{X}) \geqslant \gamma$. Set $t = 1$ (counter).
2. For $i = 1$ to $n = 10$:
   a. Let $\mathcal{P}_i$ be the set of paths containing link i.
   b. If $\max_{k\notin\mathcal{P}_i} L_k \geqslant \gamma$, then draw $X_i \sim \mathsf{Exp}(1)$.
   c. Otherwise, set $X_i = 0$ and compute $L = \max_{k\in\mathcal{P}_i} L_k$. Draw $Z \sim \mathsf{Exp}(1)$ and set $X_i = Z + \max\{\gamma - L, 0\}$.
3. Set $t = t + 1$. If $t > N$ (sample size) stop; otherwise, return to Step 2.

The estimate and relative error are comparable to the ones obtained via the multi-level approach. See also Chan (2010) for an application to a high-dimensional network.

3.2. Transformed likelihood ratio

The *transform likelihood ratio* (TLR) method (Kroese and Rubinstein, 2004) is a convenient constructing efficient importance sampling estimators.

The idea is to apply a simple *change of variable* step to the estimation problem and then to apply the CE method to find the optimal importance sampling parameter for the transformed problem. Suppose the objective is to estimate $\ell = \mathbb{E}H(\mathbf{X})$. The main step is to write $\mathbf{X}$ as a function of another random vector, say as

$$\mathbf{X} = G(\mathbf{Z}). \tag{13}$$

If we define

$$\widetilde{H}(\mathbf{Z}) = H(G(\mathbf{Z})),$$

then estimating ℓ is equivalent to estimating

$$\ell = \mathbb{E}[\widetilde{H}(\mathbf{Z})]. \tag{14}$$

As an example, consider the one-dimensional case were $X \sim \mathsf{Weib}(\alpha,\lambda)$. Since X has the same distribution as $Z^{1/\alpha}/\lambda$, where $Z \sim \mathsf{Exp}(1)$, we have $\widetilde{H}(Z) = H(\lambda^{-1}Z^{1/\alpha})$ and $\ell = \mathbb{E}[H(\lambda^{-1}Z^{1/\alpha})]$.

To apply the cross-entropy method, assume that $\mathbf{Z}$ has a density $h(\mathbf{z};\boldsymbol{\theta})$ in some class of densities $\{h(\mathbf{z};\boldsymbol{\eta})\}$. Then we can seek to estimate ℓ efficiently via importance sampling. In particular, by analogy to (3), we obtain the following TLR estimator:

$$\hat{\ell} = \frac{1}{N}\sum_{k=1}^{N} \widetilde{H}(\mathbf{Z}_k)\,\widetilde{W}(\mathbf{Z}_k;\boldsymbol{\theta},\boldsymbol{\eta}), \tag{15}$$

where

$$\widetilde{W}(\mathbf{Z}_k;\boldsymbol{\theta},\boldsymbol{\eta}) = \frac{h(\mathbf{Z}_k;\boldsymbol{\theta})}{h(\mathbf{Z}_k;\boldsymbol{\eta})}$$

and $\mathbf{Z}_k \sim h(\mathbf{z};\boldsymbol{\eta})$. As an example, consider again the $\mathsf{Weib}(\alpha,\lambda)$ case. Using the transform $X = Z^{1/\alpha}/\lambda$, we could apply importance sampling to $Z \sim \mathsf{Exp}(1)$, using an $\mathsf{Exp}(\eta)$ class of distributions. Thus, $h(z;\eta) = \eta\,\mathrm{e}^{-\eta z}$ is the importance sampling pdf, with $\eta = \theta = 1$ as the nominal parameter. Hence, in this case, $\hat{\ell}$ in (15) reduces to

$$\hat{\ell} = \frac{1}{N}\sum_{k=1}^{N} \widetilde{H}\left(\lambda^{-1}Z_k^{1/\alpha}\right)\widetilde{W}(Z_k;\theta,\eta), \tag{16}$$

with

$$\widetilde{W}(Z_k;\theta,\eta) = \frac{h(Z_k;\theta)}{h(Z_k;\eta)} = \frac{\theta\,\mathrm{e}^{-\theta Z_k}}{\eta\,\mathrm{e}^{-\eta Z_k}}$$

and $Z_k \sim \mathsf{Exp}(\eta)$.

To find the optimal parameter vector $\boldsymbol{\eta}^*$ of the TLR estimator (15) we can solve, by analogy to (5), the following CE program:

$$\boldsymbol{\eta}^* = \operatorname*{argmax}_{\boldsymbol{\eta}} \mathbb{E}_{\boldsymbol{\tau}}\left[\widetilde{H}(\mathbf{Z})\widetilde{W}(\mathbf{Z};\boldsymbol{\theta},\boldsymbol{\tau})\ln h(\mathbf{Z};\boldsymbol{\eta})\right] \tag{17}$$

and similarly for the stochastic counterpart of (17).

To obtain simple updating formulas one would typically choose the distribution of $\mathbf{Z}$ from an exponential family of distributions, as is explained in Section 2.3. Below we present the TLR algorithm for estimating $\ell = \mathbb{E}_f[H(\mathbf{X})]$, assuming that $\mathbf{X}$ is a random vector with independent, continuously distributed components.

Algorithm 3.3 (TLR Method)

1. For a given random vector $\mathbf{X}$, find a transformation G such that $\mathbf{X} = G(\mathbf{Z})$, with $\mathbf{Z} \sim h(\mathbf{z};\boldsymbol{\theta})$. For example, take $\mathbf{Z}$ with all components being iid and distributed according to an exponential family (e.g., $\mathsf{Exp}(1)$).
2. Generate a random sample $\mathbf{Z}_1,\ldots,\mathbf{Z}_N$ from $h(\cdot;\boldsymbol{\tau})$.
3. Solve the stochastic counterpart of the program (17). Iterate if necessary. Denote the solution by $\hat{\boldsymbol{\eta}}$.

4. Generate a (larger) random sample $\mathbf{Z}_1, \ldots, \mathbf{Z}_{N_1}$ from $h(\cdot; \hat{\boldsymbol{\eta}})$ and estimate $\ell = \mathbb{E}[H(G(\mathbf{Z}))]$ via the TLR estimator (15), taking $\boldsymbol{\eta} = \hat{\boldsymbol{\eta}}$.

The advantage of the TLR method is its universality and its ability to avoid the computational burden while directly delivering the analytical solution of the stochastic counterpart of the program (17).

Example 3.3 (Elliptical Distributions) A random vector $\mathbf{X} = (X_1, \ldots, X_n)$ is said to have an *elliptical distribution* with location vector $\boldsymbol{\mu}$, dispersion matrix Σ, and radial cdf F, written $\mathbf{X} \sim \mathsf{Ellipt}(\boldsymbol{\mu}, \Sigma, F)$, if $\mathbf{X}$ can be written in the form

$$\mathbf{X} = \boldsymbol{\mu} + RB\mathbf{Y}, \tag{18}$$

where $\mathbf{Y}$ is a uniform vector on the n-dimensional sphere, R is a random variable with cdf F, independent of $\mathbf{Y}$, and B is a matrix such that $BB^\top = \Sigma$. Note that a random vector that is uniformly distributed on the n-dimensional sphere can be obtained by normalizing an n-dimensional standard normal random vector: $\mathbf{Y} = \mathbf{Z}/\|\mathbf{Z}\|$, with $\mathbf{Z} \sim \mathsf{N}(\mathbf{0}, I_n)$; see, for example, Kroese et al. (2011); Blanchet and Rojas-Nandayapa (in press) consider efficient simulation procedures for estimating probabilities of the form

$$\ell = \mathbb{P}(\mathrm{e}^{X_1} + \cdots + \mathrm{e}^{X_n} \geqslant \gamma), \tag{19}$$

where $\mathbf{X} \sim \mathsf{Ellipt}(\boldsymbol{\mu}, \Sigma, F)$. We show how such quantities can be quickly computed via the TLR method. Defining

$$S(R, \mathbf{Y}) = \mathrm{e}^{X_1} + \cdots + \mathrm{e}^{X_n},$$

where the $\{X_i\}$ are related to the $\{Y_i\}$ and R via (18), the idea is to write (19) as

$$\ell = \mathbb{P}(S(R, \mathbf{Y}) \geqslant \gamma)$$

and to estimate the latter by performing importance sampling on the distribution of R only. Suppose, for simplicity, that $R \sim \mathsf{Exp}(1)$ and that the importance sampling distribution is $\mathsf{Exp}(1/\hat{v}_T)$ where $\hat{v}_T$ is obtained at the last step of a CE procedure. This means that ℓ is estimated as

$$\hat{\ell} = \frac{1}{N_1} \sum_{k=1}^{N_1} \mathrm{I}_{\{S(R_k, \mathbf{Y}_k) \geqslant \gamma\}} W_k, \tag{20}$$

where $W_k = \hat{v}_T \, \mathrm{e}^{-(1-1/\hat{v}_T)R_k}$. Because $\{\mathsf{Exp}(1/v), v > 0\}$ is an exponential family parameterized by its mean, the optimal reference parameter at the t-th iteration of the CE algorithm is given by

$$\hat{v}_t = \frac{\sum_{k=1}^{N} \mathrm{I}_{\{S(R_k, \mathbf{Y}_k) \geqslant \hat{\gamma}_t\}} W_k R_{ki}}{\sum_{k=1}^{N} \mathrm{I}_{\{S(R_k, \mathbf{Y}_k) \geqslant \hat{\gamma}_t\}} W_k}, \tag{21}$$

where $W_k = \hat{v}_{t-1} \mathrm{e}^{-(1-1/\hat{v}_{t-1})R_k}$. This leads to the following procedure.

Algorithm 3.4

1. Set $\hat{v}_0 = u = 1$ and $t = 1$.
2. Generate $R_1, \ldots, R_N \sim_{\text{iid}} \mathsf{Exp}(\hat{v}_{t-1})$ and (independently) $\mathbf{Z}_k \sim \mathsf{N}(\mathbf{0}, \mathrm{I}_n), k = 1, \ldots, N$. Let $\mathbf{Y}_k = \mathbf{Z}_k / \|\mathbf{Z}_k\|, k = 1, \ldots, N$.
3. Set $\mathbf{X}_k = \boldsymbol{\mu} + R_k B \mathbf{Y}_k$, calculate the performances $S(R_k, \mathbf{Y}_k)$ and order these from smallest to largest: $S_{(1)} \leqslant \cdots \leqslant S_{(N)}$. Let $\hat{\gamma}_t$ be the sample $(1-\varrho)$-quantile of performances. If $\hat{\gamma}_t > \gamma$, reset $\hat{\gamma}_t$ to γ.
4. Update $\hat{v}_t$ via (21), using the **same** $R_1, \ldots, R_N$ as obtained in Step 2.
5. If $\hat{\gamma}_t < \gamma$, set $t = t + 1$ and reiterate from Step 2; otherwise, proceed with Step 6.
6. Let T be the final iteration counter. Generate $R_1, \ldots, R_{N_1} \sim_{\text{iid}} \mathsf{Exp}(\hat{v}_T)$ and $\mathbf{Y}_1, \ldots, \mathbf{Y}_{N_1}$, and estimate ℓ via importance sampling as in (20).

As a numerical example, consider the 10-dimensional case with $\boldsymbol{\mu} = \mathbf{0}, \Sigma = I_{10}$ (identity), $R \sim \mathsf{Exp}(1)$, and $\gamma = 3 \cdot 10^3$. Using a sample size of $N = 10^3$ the following CE parameter was found after three iterations: $\hat{v}_3 = 12.05$. Using importance sampling with a sample size of 10^6 the rare-event probability was estimated as $3.91 \cdot 10^{-6}$ with an estimated relative error of 1.2%. For $\gamma = 3 \cdot 10^5$, using the same sample sizes, the probability was estimated (after five iterations) as $9.3 \cdot 10^{-9}$ with an estimated relative error of 2.6%.

3.3. Root finding

In many applications one needs to estimate, for given ℓ, the *root*, γ, of the nonlinear equation

$$\mathbb{P}_{\mathbf{u}}(S(\mathbf{X}) \geqslant \gamma) = \mathbb{E}_{\mathbf{u}}[\mathrm{I}_{\{S(\mathbf{X}) \geqslant \gamma\}}] = \ell \tag{22}$$

rather than estimate ℓ itself. We call such a problem a *root-finding* problem.

One can obtain γ using the CE method, by finding a good reference vector $\hat{\mathbf{v}}_T$ such that γ can be estimated accurately as the smallest number $\hat{\gamma}$ such that

$$\frac{1}{N_1} \sum_{k=1}^{N_1} \mathrm{I}_{\{S(\mathbf{X}_k) \geqslant \hat{\gamma}\}} W(\mathbf{X}_k; \mathbf{u}, \hat{\mathbf{v}}_T) \leqslant \ell. \tag{23}$$

To find such a reference vector $\hat{\mathbf{v}}_T$ one can simply modify the multi-level CE Algorithm 2.2 as follows.

Algorithm 3.5 (Root-Finding Algorithm)

1. Define $\hat{\mathbf{v}}_0 = \mathbf{u}, N^{\mathrm{e}} = \lceil (1-\varrho) N \rceil$. Set $t = 1$.
2. Generate a random sample $\mathbf{X}_1, \ldots, \mathbf{X}_N$ from the density $f(\cdot; \hat{\mathbf{v}}_{t-1})$.
3. Calculate the performances $S(\mathbf{X}_1), \ldots, S(\mathbf{X}_N)$. Order the performances from smallest to largest: $S_{(1)} \leqslant \cdots \leqslant S_{(N)}$. Let $\hat{\gamma}_t = S_{(N^{\mathrm{e}})}$.
4. Calculate $\hat{\ell}_t = \max\{\ell, \frac{1}{N} \sum_{k=1}^{N} I_{\{S(\mathbf{X}_k) \geqslant \hat{\gamma}_t\}} W(\mathbf{X}_k; \mathbf{u}, \hat{\mathbf{v}}_{t-1})\}$.
5. Solve the stochastic program (6) with $\mathbf{w} = \hat{\mathbf{v}}_{t-1}$, using the **same** sample $\mathbf{X}_1, \ldots, \mathbf{X}_N$. Denote the solution by $\hat{\mathbf{v}}_t$.
6. If $\hat{\ell}_t = \ell$, proceed to Step 7; otherwise, let $t = t + 1$ and reiterate from Step 2.

Table 2
Convergence of the sequence $\{(\hat{\ell}_t, \hat{\mathbf{v}}_t)\}$

t	$\hat{\ell}_t$	$\hat{\mathbf{v}}_t$									
0	–	1	1	1	1	1	1	1	1	1	1
1	0.1	1.46	1.44	1.77	1.44	1.44	1.44	1.43	1.16	1.72	1.70
2	$4.28 \cdot 10^{-3}$	1.74	1.83	2.54	1.79	1.81	1.75	1.78	1.19	2.39	2.45
3	$1.75 \cdot 10^{-4}$	2.04	1.99	3.25	2.10	2.01	2.00	2.24	1.07	3.11	3.11
4	$1.00 \cdot 10^{-5}$	2.42	2.28	3.92	2.18	2.28	2.15	2.68	1.03	3.54	4.12

7. Estimate γ via the right-hand side of (23), using a sample $\mathbf{X}_1, \ldots, \mathbf{X}_{N_1} \sim f(\cdot; \hat{\mathbf{v}}_T)$, where T is the final iteration number.

Example 3.4 (Root Finding for the Stochastic Activity Network) Consider the stochastic activity network in Example 2.2. Suppose we wish to estimate for which γ

$$\mathbb{P}(S(\mathbf{X}) \geqslant \gamma) = 10^{-5}.$$

Table 2 shows a typical outcome of Algorithm 3.5, with a sample size of $N = 10^5$ and $\varrho = 0.1$. A typical estimate of γ using a sample size of $N_1 = 10^6$ is 18.08 with an estimated relative error of 0.1%.

Acknowledgment

This work was supported by the Australian Research Council under Grant No. DP0985177.

References

Asmussen, S., Kroese, D.P., Rubinstein, R.Y., 2005. Heavy tails, importance sampling and cross-entropy. Stoch. Models 210 (1), 57–76.

Blanchet, J.H., Rojas-Nandayapa, L., 2011. Efficient simulation of tail probabilities of sums of dependent random variables. J. Appl. Probab. 48A, 147–164.

Chan, J.C.C., 2010. Advanced Monte Carlo methods with applications in finance. PhD Thesis. University of Queensland.

Chan, J.C.C., Kroese, D.P., 2010. Efficient estimation of large portfolio loss probabilities in t-copula models. Eur. J. Oper. Res. 2050 (2), 361–367.

Chan, J.C.C., Kroese, D.P., 2011. Rare-event probability estimation with conditional Monte Carlo. Ann. Oper. Res. 189 (1), 155–165.

Chan, J.C.C., Kroese, D.P., 2012. Improved cross-entropy method for estimation. Stat. Comput. 22 (5), 1031–1040.

Chan, J.C.C, Glynn P.W., Kroese, D.P., 2011. A comparison of cross-entropy and variance minimization strategies. J. Appl. Prob. 48A, 183–194.

de Boer, P.T., 2000. Analysis and efficient simulation of queueing models of telecommunication systems. PhD Thesis. University of Twente.

de Boer, P.T., Kroese, D.P., Rubinstein, R.Y., 2004. A fast cross-entropy method for estimating buffer overflows in queueing networks. Manag. Sci. 500 (7), 883–895.

de Boer, P.T., Kroese, D.P., Mannor, S., Rubinstein, R.Y., 2005. A tutorial on the cross-entropy method. Ann. Oper. Res. 1340 (1), 19–67.

Evans, G.E., 2009. Parallel and sequential Monte Carlo methods with applications. PhD Thesis. The University of Queensland, Brisbane.

Evans, G.E., Keith, J.M., Kroese, D.P., 2007. Parallel cross-entropy optimization. In: Proceedings of the 2007 Winter Simulation Conference, Washington, DC, pp. 2196–2202.

Homem-de-Mello, T., 2007. A study on the cross-entropy method for rare event probability estimation. INFORMS J. Comput. 190 (3), 381–394.

Hui, K-P., Bean, N., Kraetzl, M., Kroese, D.P., 2005. The cross-entropy method for network reliability estimation. Ann. Oper. Res. 134, 101–118.

Kroese, D.P., Hui, K.-P., 2006. In: Computational Intelligence in Reliability Engineering, chapter 3: Applications of the Cross-Entropy Method in Reliability. Springer-Verlag, New York.

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

Kroese, D.P., Taimre, T., Botev, Z.I., 2011. Handbook of Monte Carlo Methods. John Wiley & Sons, New York.

Rao, C.R., 2010. Entropy and cross entropy: characterizations and applications. In: Alladi, K., Klauder, J., Rao, C.R. (Eds.), The Legacy of Alladi Ramakrishnan in the Mathematical Sciences. Springer-Verlag, New York, pp. 359–367.

Ridder, A., 2005. Importance sampling simulations of Markovian reliability systems using cross-entropy. Ann. Oper. Res. 1340 (1), 119–136.

Rubinstein, R.Y., 1997. Optimization of computer simulation models with rare events. Eur. J. Oper. Res. 990 (1), 89–112.

Rubinstein, R.Y., 1999. The cross-entropy method for combinatorial and continuous optimization. Methodol. Comput. Appl. Probab. 10 (2), 127–190.

Rubinstein, R.Y., 2001. Combinatorial optimization, cross-entropy, ants and rare events. In: Uryasev, S., Pardalos, P.M. (Eds.), Stochastic Optimization: Algorithms and Applications, Kluwer, Dordrecht, pp. 304–358.

Rubinstein, R.Y., Kroese, D.P., 2004. The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization. Monte Carlo Simulation and Machine Learning. Springer-Verlag, New York.

Rubinstein, R.Y., Kroese, D.P., 2007. Simulation and the Monte Carlo Method, second ed. John Wiley & Sons, New York.

Taimre, T., 2009. Advances in cross-entropy methods. PhD Thesis. The University of Queensland, Brisbane.