AN EXPLICIT CROSS ENTROPY SCHEME FOR MIXTURES

HUI WANG† AND XIANG ZHOU‡

Abstract. The key issue in importance sampling is the choice of the alternative sampling distribution, which is often chosen from the exponential tilt family of the underlying distribution. However, when the problem exhibits certain kind of nonconvexity, it is very likely that a single exponential change of measure will never attain asymptotic optimality and may lead to erroneous estimates. In this paper we introduce an explicit iterative scheme which combines the traditional cross-entropy method and the EM algorithm to find an efficient alternative sampling distribution in the form of mixtures. We also study the applications of this scheme to option price estimation.

Key words. importance sampling, cross-entropy, mixture, option pricing

AMS subject classifications. 00A72, 91G60

1. Introduction. Importance sampling is a variance reduction technique in Monte Carlo simulation. It is particularly powerful in the context of rare event simulation [1, 2, 10, 15, 19, 21, 22] and has numerous applications in computational finance [12, 13, 14, 23]. The main idea of importance sampling is to generate samples from an alternative sampling distribution and multiply the outcome with the appropriate likelihood ratio in order to obtain unbiased estimates. To be more concrete, consider a generic problem of estimating

\[ \mu = \mathbb{E}[V(X)], \]

where \( X \) is an \( \mathbb{R}^d \)-valued random variable defined on some probability space \((\Omega, \mathcal{F}, P)\) and \( V : \mathbb{R}^d \rightarrow \mathbb{R} \) is some given nonnegative function. For simplicity, further assume that \( X \) has a density function \( f \). Observe that for any positive density function \( g \),

\[ \mu = \int_{\mathbb{R}^d} V(x)f(x) dx = \int_{\mathbb{R}^d} V(x)f(x) g(x)dx. \]  

(1.1)

Importance sampling estimates \( \mu \) by the sample average of independent identically distributed (iid) copies of

\[ \hat{\mu} = V(X)\frac{f(X)}{g(X)}, \]

where, abusing notation, \( X \) is a representative sample from the new density function \( g \). The estimate \( \hat{\mu} \) is clearly unbiased, thanks to (1.1).

The efficiency of the importance sampling scheme depends on the alternative probability distribution \( g \). A popular class of alternative probability measures, suggested by the large deviation analysis, is the so-called “exponentially tilted distribution”, where

\[ g(x) = e^{\langle \theta, x \rangle - H(\theta)}, \quad H(\theta) = \log \int_{\mathbb{R}^d} e^{\langle \theta, x \rangle} f(x) dx, \]

*This research is supported in part by the National Science Foundation (NSF-DMS-1008331).
†Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912, USA (huiwang@cfm.brown.edu).
‡Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong, P.R. China (xiang.zhou@cityu.edu.hk).

arXiv:1305.3226v1 [math.PR] 14 May 2013
for some $\theta \in \mathbb{R}^d$. The cross-entropy method, invented by Reuven Rubinstein [16, 17, 18], is a simulation based technique for selecting the tilting parameter $\theta$. It is easy to implement and does not require much additional computational cost. The implicit assumption for the resulting scheme to work well is that the exponential tilt family contains a member which serves as a good sampling distribution.

However, it is well known that using a single exponentially tilted distribution will lead to suboptimality and erroneous estimates [11, 6] except in very simplistic settings. Therefore, it is in general fruitless to look for a good alternative sampling distribution within the exponential tilt family. One has to enlarge the class to have a chance of finding a good sampling distribution.

A possible solution to a general importance sampling problem is to use the game/subsolution framework developed in [8, 9], where the resulting alternative sampling distributions are usually state-dependent. Even though it can be shown that such schemes are optimal in an appropriate sense, the implementation requires the user to construct suitable subsolutions to certain partial differential equations, which is itself a highly nontrivial task for many practical applications with nonhomogeneous dynamics.

In this paper we consider a different approach, where the class of alternative sampling distributions consists of mixtures of exponentially tilted distributions. In many practical applications, the class of mixtures is often sufficient due to the special structures associated with the problems under consideration [5, 20]. Using cross-entropy to determine a mixture sampling distribution has been considered in [4, 3]. But the approaches there are very different. In [4] the authors mainly consider rare event probabilities and the weights are estimated from the samples by dividing the set of interest. It is not clear whether such weights are optimal. Moreover, their approach requires samples from approximate zero-variance change of measure, which is achieved from MCMC sampling and is itself difficult. [3] is concerned with estimating the tail probabilities of the sum of iid heavy-tailed random variables. There the components of the mixture are given and only the weights are optimized through the classical iterative cross-entropy scheme. The main purpose of [3] is not developing general schemes but rather analyzing the efficiency of mixture sampling distributions in the context of heavy-tailed distributions. Our scheme is different in that we combine cross-entropy with the classical Expectation-Maximization (EM) algorithm [7] to produce a good mixture sampling distribution for general models. Our scheme leads to explicit iterations to find both the optimal weight and the optimal exponential tilting parameter for each component of the mixture.

**Comments on notation:** Here we clarify some of the convention and notation used throughout the paper.

1. To ease exposition, in all the discussions we assume that the underlying distribution admits a density. The extension to the general distributions is straightforward.

2. The expected value with respect to the original distribution is denoted by $E$. Given an alternative density function $g$, the notation

$$E_g[V(X)]$$

denotes the expected value when $X$ has density $g$. If the alternative density belongs to some parametrized family, say $\{f_\theta\}$, then the notation is simplified
to

$$E_\theta [V(X)] = E_{f_\theta} [V(X)],$$

provided there is no ambiguity.

3. Suppose \( X \) has a density \( f \) and one wishes to estimate \( E[V(X)] \). The alternative sampling distribution \( g \) must satisfy the "absolute continuity" condition, that is, \( V(x)f(x) = 0 \) whenever \( g(x) = 0 \). Throughout the paper this condition will be implicitly imposed.

4. \( \langle \cdot , \cdot \rangle \) denotes the usual Euclidean inner product of two vectors. If the dimension is one, then the inner product is just the ordinary product.

5. The notation

$$\frac{\partial}{\partial \theta}$$

denotes the partial derivative with respect to the argument \( \theta \). If \( \theta \) happens to be a vector, then it denotes the gradient with respect to \( \theta \).

6. The \( d \times d \) identity matrix is denoted by \( I_d \).

7. Inequalities between any two vectors hold in the componentwise sense.

2. Basics of Cross-Entropy. In this section we describe the main idea of the cross-entropy method. A detailed account can be found in [18]. To illustrate, consider the generic problem of estimating the expected value

$$\mu = E[V(X)],$$

where \( V \) is a nonnegative function and \( X \) has density \( f \). Assume that the alternative sampling distribution is restricted to a parameterized family, say \( \{f_\theta\} \). A particularly popular choice is the exponential tilt family described in the Introduction.

It is well known that there exists a zero-variance alternative sampling distribution \( f^* \) defined by

$$f^*(x) = \frac{1}{\mu} V(x)f(x).$$

In other words, under \( f^* \), the estimate is the constant \( \mu \) and has zero variance. Even though such a sampling distribution is impractical as it requires the knowledge of \( \mu \), it leads to the heuristic principle that an alternative probability distribution "close" to \( f^* \) should be a good choice for importance sampling. The cross-entropy method aims to solve for the alternative probability distribution \( f_\theta \) that is closest to \( f^* \) in the sense of Kullback-Leibler cross-entropy or relative entropy, defined by

$$R(f^*||f_\theta) = \int f^*(x) \log \frac{f^*(x)}{f_\theta(x)} \, dx.$$  

That is, the cross-entropy method chooses the minimizing probability density to the minimization problem

$$\min_\theta R(f^*||f_\theta)$$

as the alternative sampling distribution for importance sampling. Plugging in formula [2.1], it follows that

$$R(f^*||f_\theta) = \int f^*(x) \log \frac{f^*(x)}{f(x)} \, dx + \int f^*(x) \log \frac{f(x)}{f_\theta(x)} \, dx$$
\[ R(f^* \| f) + \frac{1}{\mu} \int V(x) \log \frac{f(x)}{f_\theta(x)} \cdot f(x) \, dx \]

\[ = R(f^* \| f) + \frac{1}{\mu} E[V(X) \log f(X)] - \frac{1}{\mu} E[V(X) \log f_\theta(X)], \]

Since the first term, the second term, and \( \mu \) are all independent of \( \theta \), the minimization problem (2.2) is equivalent to the maximization problem

\[ \max_\theta E[V(X) \log f_\theta(X)]. \] (2.3)

The maximizing \( \theta \), under mild regularity conditions, solves the equation

\[ 0 = \frac{\partial}{\partial \theta} E[V(X) \log f_\theta(X)] = E[V(X) \frac{\partial}{\partial \theta} \log f_\theta(X)]. \] (2.4)

Even though there is no explicit solution to the equation in general, the cross-entropy method produces a simple simulation based algorithm to approximate the solution. To be more precise, if one replaces the expected value by sample average and considers the corresponding stochastic version

\[ 0 = \frac{1}{N} \sum_{k=1}^{N} V(X_k) \frac{\partial}{\partial \theta} \log f_\theta(X_k), \]

where \( \{X_k\} \) are iid copies of \( X \), then it is often possible to obtain a solution in closed form; see [18, Appendix A]. To summarize, the basic cross-entropy method generates \( N \) iid pilot samples \( X_1, \ldots, X_N \) from the original distribution and compute \( \theta \) from equation (2.4). Once \( \theta \) is obtained, \( \mu = E[V(X)] \) is estimated by importance sampling using the alternative sampling density \( f_\theta \).

The basic cross-entropy algorithm can easily extends to a more general, iterative procedure for solving the minimization problem (2.3). Every iteration in the general scheme involves two phases. In the \((i+1)\)-th iteration, one first generates iid pilot samples \( X_1, \ldots, X_N \) from density \( f_\theta \) with \( \theta = \hat{\theta}_i \) being the current candidate of the tilting parameter. The tilting parameter \( \hat{\theta}_i \) is then updated to \( \hat{\theta}_{i+1} \) based on these samples. As in the basic algorithm, \( \hat{\theta}_{i+1} \) often admits closed form formulas. If \( \hat{\theta}^* \) is the tilting parameter from the final iteration, then \( f_{\hat{\theta}^*} \) is used as the alternative sampling density to estimate \( \mu \).

The iterative cross-entropy algorithm is based on the following observation. Fixing an arbitrary tilting parameter say \( \theta^0 \), we can rewrite the expected value in (2.3) as

\[ E_{\theta^0} [V(X) \ell_{\theta^0}(X) \log f_\theta(X)], \] (2.5)

where for any \( \theta \),

\[ \ell_{\theta}(x) = \frac{f(x)}{f_\theta(x)}, \]

denotes the likelihood ratio. Consequently, the maximizer to (2.3) satisfies the equation

\[ 0 = E_{\theta^0} \left[ V(X) \ell_{\theta^0}(X) \frac{\partial}{\partial \theta} \log f_\theta(X) \right]. \]
As before, we replace the expected value by sample average and solve the equation

\[ 0 = \frac{1}{N} \sum_{k=1}^{N} V(X_k) \ell_{\hat{\theta}}(X_k) \frac{\partial}{\partial \theta} \log f_\theta(X_k), \]

where \( X_1, \ldots, X_N \) are iid pilot samples from the probability density \( f_{\hat{\theta}} \). This leads to the following updating rule for \( \hat{\theta} \).

**The updating rule of \( \hat{\theta} \).** Suppose that \( \hat{\theta}^i \) is the value of the tilting parameter at the end of the \( i \)-th iteration. Then \( \hat{\theta}^{i+1} \) is set to be the solution to the equation

\[ 0 = \frac{1}{N} \sum_{k=1}^{N} V(X_k) \ell_{\hat{\theta}^i}(X_k) \frac{\partial}{\partial \theta} \log f_\theta(X_k), \]

where \( X_1, \ldots, X_N \) are iid pilot samples from the probability density \( f_{\hat{\theta}^i} \).

Equation (2.6) is of exactly the same form as the basic cross-entropy equation (2.4), except that \( V(X_k) \) is replaced by \( V(X_k) \ell_{\hat{\theta}^i}(X_k) \), thus often explicitly solvable as well. In particular, we have the following result concerning the normal distributions, in which case an exponentially tilted distribution amounts to a shift in the mean. Its proof is trivial and thus omitted.

**Lemma 2.1.** Suppose that \( f \) is the density of \( N(0, I_d) \) and \( f_\theta \) denotes the density of \( N(\theta, I_d) \) for \( \theta \in \mathbb{R}^d \). Then the solution to (2.6) is simply

\[ \hat{\theta}^{i+1} = \frac{\sum_{k=1}^{N} V(X_k) e^{-\langle \hat{\theta}^i, X_k \rangle} X_k}{\sum_{k=1}^{N} V(X_k) e^{-\langle \hat{\theta}^i, X_k \rangle}}, \]

where \( X_1, \ldots, X_N \) are iid pilot samples from \( N(\hat{\theta}^i, I_d) \).

**Remark 1.** Detailed discussion of the initialization of \( \hat{\theta}^0 \) can be found in [18, Chapter 8]. We will also touch upon it later in the paper. It is worth pointing out that if the initial tilting parameter \( \hat{\theta}^0 \) is chosen such that \( f_{\hat{\theta}^0} = f \), then the first step of iterative cross entropy algorithm is exactly the basic cross entropy algorithm since \( \ell_{\hat{\theta}^0}(x) = 1 \) for all \( x \).

**3. Cross Entropy for Mixture.** It is often the case that a single exponential change of measure will not be sufficient for constructing efficient importance sampling algorithms. This was first observed in [11] and has motivated extensive investigations afterwards, which have led to schemes such as dynamic importance sampling [8, 9] and Lyapunov function method [2]. As powerful as these methods are, they are more accessible when the underlying dynamics are piecewise homogeneous (such as queueing models), for it is easier to explore the structure of some related partial differential equations. In financial models, however, due to the nonlinearity commonly associated with diffusive dynamics, it is very difficult to extract information from such partial differential equations.

In this section, we consider an alternative method based on mixtures, which take the following form. Let \( \{f_\alpha(x) : x \in \mathbb{R}^d\} \) be a family of density functions. Given a positive integer \( m \), let \( \theta = (w_1, \ldots, w_m; \alpha_1, \ldots, \alpha_m) \), where \((w_1, \ldots, w_m)\) is a vector of weights, that is,

\[ w_i > 0, \quad w_1 + \cdots + w_m = 1, \]
and define the mixture density

\[ h_\theta(x) = \sum_{j=1}^{m} w_j \cdot f_{\alpha_j}(x). \]

In practice, \(\{f_\alpha\}\) is usually chosen to be the family of exponentially tilted distributions, and \(\{h_\theta\}\) represents their mixtures.

Recall that we are interested in estimating the expected value \(\mu = E[V(X)]\), where \(V\) is nonnegative and \(X\) has density \(f\). Fix an arbitrary parameter \(\theta^0 = (w_1^0, \ldots, w_m^0; \alpha_1^0, \ldots, \alpha_m^0)\). As before, the cross-entropy scheme amounts to solving the maximization problem (2.5), namely,

\[ \max_{\theta} E_{\theta^0}[V(X)\ell_{\theta^0}(X) \log h_\theta(X)], \]

where for any \(\theta\)

\[ \ell_\theta(x) \triangleq \frac{f(x)}{h_\theta(x)}. \]

The stochastic version of the previous maximization problem is to maximize

\[ \frac{1}{N} \sum_{k=1}^{N} \tilde{V}(X_k) \log h_\theta(X_k), \]

where \(X_1, \ldots, X_N\) are iid samples from the mixture density \(h_{\theta^0}(x)\) and

(3.1) \[ \tilde{V}(x) \triangleq V(x)\ell_{\theta^0}(x). \]

The problem here is that this maximization problem will not have explicit solution anymore.

To resolve this issue, we resort to the idea of EM algorithm. Given any \(\theta\), define (abusing notation) the joint probability distribution function by

\[ h_\theta(x, j) \triangleq w_j \cdot f_{\alpha_j}(x), \quad \text{for} \ x \in \mathbb{R}^d \ \text{and} \ j = 1, \ldots, m. \]

Note that the marginal of \(h_\theta(x, \cdot)\) on \(x\) is exactly the mixture \(h_\theta(x)\). Now we can write, for any \(j = 1, \ldots, m\),

\[ \tilde{V}(X_k) \log h_\theta(X_k) = \tilde{V}(X_k) \log h_\theta(X_k, j) - \tilde{V}(X_k) \log \frac{h_\theta(X_k, j)}{h_\theta(X_k)} \]

\[ = \tilde{V}(X_k) \log h_\theta(X_k, j) - \tilde{V}(X_k) \log h_\theta(j|X_k), \]

where \(h_\theta(j|x) = h_\theta(x, j)/h_\theta(x)\) is the conditional probability distribution function. Since the above display is true for any possible value of \(j\), we can integrate both sides against the conditional distribution \(h_{\theta^0}(\cdot|x)\) and obtain

\[ \tilde{V}(X_k) \log h_\theta(X_k) = \sum_{j=1}^{m} \tilde{V}(X_k) \log h_\theta(X_k, j) \cdot h_{\theta^0}(j|X_k) \]

\[ - \sum_{j=1}^{m} \tilde{V}(X_k) \log h_\theta(j|X_k) \cdot h_{\theta^0}(j|X_k) \]
\[ \sum_{j=1}^{m} \tilde{V}(X_k) \log h_\theta(X_k, j) \cdot h_{\theta^0}(j|X_k) \\
- \sum_{j=1}^{m} \tilde{V}(X_k) \log h_{\theta^0}(j|X_k) \cdot h_{\theta^0}(j|X_k) \\
+ \sum_{j=1}^{m} \tilde{V}(X_k) \log \frac{h_{\theta^0}(j|X_k)}{h_\theta(X_k)} \cdot h_{\theta^0}(j|X_k). \]

Following the ideas of EM algorithm, we choose \( \theta \) that maximizes the summation (over \( k \)) of the first term on the right-hand-side, since the second term is independent of \( \theta \) and the last term is nonnegative (which is essentially a relative entropy itself and takes value zero if \( \theta = \theta^0 \)). That is, we choose \( \theta \) to maximize

\[ \sum_{k=1}^{N} \sum_{j=1}^{m} \tilde{V}(X_k) \log \frac{w_j f_{\alpha_j}(X_k)}{h_\theta(X_k)} \cdot h_{\theta^0}(j|X_k) \]

\[ = \sum_{j=1}^{m} \sum_{k=1}^{N} \tilde{V}(X_k) \log \frac{w_j f_{\alpha_j}(X_k)}{h_\theta(X_k)} \cdot h_{\theta^0}(j|X_k) \]

\[ = \sum_{j=1}^{m} \sum_{k=1}^{N} \tilde{V}(X_k) \log f_{\alpha_j}(X_k) \cdot h_{\theta^0}(j|X_k) \]

\[ + \sum_{j=1}^{m} \sum_{k=1}^{N} \tilde{V}(X_k) \log w_j \cdot h_{\theta^0}(j|X_k). \]

Note that this maximization problem is completely separated. In other words, for each \( j = 1, \ldots, m \), we choose \( \alpha_j \) to maximize

\[ (3.2) \sum_{k=1}^{N} \tilde{V}(X_k) h_{\theta^0}(j|X_k) \log f_{\alpha_j}(X_k), \]

which often admits explicit formula, and we choose \( (w_j) \) to maximize

\[ \sum_{j=1}^{m} \sum_{k=1}^{N} \tilde{V}(X_k) h_{\theta^0}(j|X_k) \log w_j, \quad \text{s.t. } w_j \geq 0, \sum_{j=1}^{m} w_j = 1, \]

which can be solved explicitly and the maximizer is

\[ (3.3) \quad w_j^* = \frac{\sum_{k=1}^{N} \tilde{V}(X_k) h_{\theta^0}(j|X_k)}{\sum_{l=1}^{m} \sum_{k=1}^{N} \tilde{V}(X_k) h_{\theta^0}(l|X_k)}. \]

Recalling the definition of \( \tilde{V} \) in (3.1), we arrive at an iterative cross-entropy scheme with the following updating rule.

**The updating rule of \( \hat{\theta} \).** Suppose that \( \hat{\theta}^i \) is the parameter for the mixture at the end of the \( i \)-th iteration. Define

\[ h_{\hat{\theta}^i}(j|x) = \frac{h_{\hat{\theta}^i}(x, j)}{h_{\hat{\theta}^i}(x)}. \]
Then $\hat{\theta}^{i+1}$ is set to be $(\hat{\omega}^{i+1}_1, \ldots, \hat{\omega}^{i+1}_m; \hat{\alpha}^{i+1}_1, \ldots, \hat{\alpha}^{i+1}_m)$, where for every $j = 1, \ldots, m$,

$$
(3.4) \quad \hat{\omega}^{i+1}_j = \frac{\sum_{k=1}^{N} V(X_k) \ell_{\hat{\theta}^i}(X_k) h_{\hat{\theta}^i}(j|X_k)}{\sum_{l=1}^{m} \sum_{k=1}^{N} V(X_k) \ell_{\hat{\theta}^i}(X_k) h_{\hat{\theta}^i}(l|X_k)},
$$

and $\hat{\alpha}^{i+1}_j$ is the solution to the equation

$$
(3.5) \quad 0 = \frac{1}{N} \sum_{k=1}^{N} V(X_k) \ell_{\hat{\theta}^i}(X_k) h_{\hat{\theta}^i}(j|X_k) \frac{\partial}{\partial \hat{\alpha}_j} \log f_{\alpha}(X_k),
$$

where $X_1, \ldots, X_N$ are iid pilot samples from the mixture density $h_{\hat{\theta}^i}$.

Equation (3.5) is of the same form as (2.6) and often admits explicit formula, especially when $\{f_\alpha\}$ is chosen to be the family of exponentially tilted distributions. The most interesting case for us is concerned with the normal distribution, where we have a result similar to Lemma 2.1.

**Lemma 3.1.** Suppose that $f$ is the density of $N(0, I_d)$ and $f_{\alpha}$ denotes the density of $N(\alpha, I_d)$ for $\alpha \in \mathbb{R}^d$. Then the solution to (3.5) is simply

$$
\hat{\alpha}^{i+1}_j = \frac{\sum_{k=1}^{N} V(X_k) \ell_{\hat{\theta}^i}(X_k) h_{\hat{\theta}^i}(j|X_k) \cdot X_k}{\sum_{k=1}^{N} V(X_k) \ell_{\hat{\theta}^i}(X_k) h_{\hat{\theta}^i}(j|X_k)},
$$

for each $j = 1, \ldots, m$.

**4. Comments on Initialization.** The initialization of the cross-entropy method involves the specification of $m$ and $\hat{\theta}^0 = (\hat{\omega}^0_1, \ldots, \hat{\omega}^0_m; \hat{\alpha}^0_1, \ldots, \hat{\alpha}^0_m)$, that is, the number of components in the mixture, as well as the respective weight and initial tilting parameter for each component. Note that if two initial tilting parameters (say $\hat{\alpha}^0_1$ and $\hat{\alpha}^0_2$) are set to be the same, then $\hat{\alpha}^0_1$ and $\hat{\alpha}^0_2$ will remain the same in each iteration, which essentially amounts to using fewer number of components in the mixture. Therefore, without loss of generality, we require that $\hat{\alpha}^0_1, \ldots, \hat{\alpha}^0_m$ be all different.

The choice of $m$ usually depends on the structure of the problem. A rule of thumb or general guidance in many financial applications is that if the function $V$ only takes positive values on the union of a collection of convex sets such as

$$\bigcup_{j=1}^{m'} A_j,$$

then it is often beneficial to set $m = m'$. We should illustrate this from numerical examples in the next section. Given $m$, the choice of the initial weights $(\hat{\omega}^0_1, \ldots, \hat{\omega}^0_m)$ seems to be most direct – setting $\hat{\omega}^0_1 = \cdots = \hat{\omega}^0_m = 1/m$ is usually a good strategy.

In the following discussion, we will focus on different approaches to initialize $(\hat{\alpha}^0_1, \ldots, \hat{\alpha}^0_m)$.

**4.1. Initialization by perturbation.** In iterative cross-entropy schemes without mixture, choice of the initial parameter can be rather flexible. A common practice is to let it be the one that corresponds to the original distribution. An immediate extension of this strategy to schemes with mixture is to combine it with random perturbation, since we require the tilting parameters to be different. For example, if a tilting parameter $\beta$ is such that

$$f_{\beta} = f,$$
then one can let
\[ \alpha_j^0 = \beta + \varepsilon_j, \quad j = 1, \ldots, m, \]
where \( \varepsilon_j \)'s are some small random perturbations.

For illustration, let us consider a very simple problem of estimating the probability of
\[
P\{X \geq a \text{ or } X \leq b\} = P\{X \in [a, \infty) \cup (-\infty, b]\},
\]
where \( X \) is a standard normal random variable and \( b < 0 < a \). In this case, \( f \) is the density of \( N(0, 1) \) and the family of exponentially tilted distributions consists of
\[ f_\alpha = \text{density of } N(\alpha, 1), \quad \alpha \in \mathbb{R}. \]

Since the target set is the union of two convex sets, we let \( m = 2 \) and \( \hat{\alpha}_1^0 = \hat{\alpha}_2^0 = 1/2 \).

As for the initial tilting parameter, we simply let
\[ \hat{\alpha}_1^0 = 0, \quad \hat{\alpha}_2^0 = -0.1. \]

The value of \( \alpha_0^2 \) is not essential. Setting \( \hat{\alpha}_2^0 \) to other similar small values will not alter the discussion below.

Table 1 is the simulation results under various parameter settings, using \( N = 20000 \) (pilot sample size), \( \text{IT}_\text{NUM} = 5 \) (number of cross-entropy iterations), and \( n = 1000000 \) (sample size for importance sampling). For comparison, plain Monte Carlo simulations with the same sample size are also performed and the entry “Var Ratio” is the ratio of the empirical variance of the plain Monte Carlo estimate to that of the cross-entropy estimate. The larger this ratio, the more significant the variance reduction from the cross-entropy algorithm. The entries \((w_1, w_2)\) and \((\alpha_1, \alpha_2)\) are the final weights and tilting parameters from the iterative cross-entropy scheme.

| \(a, b\) | True Value | CE Est | Var Ratio | \((w_1, w_2)\) | \((\alpha_1, \alpha_2)\) |
|---|---|---|---|---|---|
| \(\{1, -1.5\}\) | 0.2255 | 0.2254 | 2.4 | \((0.69, 0.31)\) | \((1.52, -1.72)\) | A |
| \(\{2, -2.5\}\) | 0.0290 | 0.0289 | 13.9 | \((0.79, 0.21)\) | \((2.37, -2.83)\) | B |
| \(\{2, -3\}\) | 0.0241 | 0.0227 | 23.6 | \((0.79, 0.20)\) | \((2.37, 2.37)\) | C |

Table 1. Estimating \( P\{X \geq a \text{ or } X \leq b\} \)

From this numerical experiment, one can say that this naive assignment of initial parameters works in very limited situations. For instance, if \( a = 1 \) and \( b = -1.5 \) (A), the cross-entropy algorithm consistently yields good performance. However, when \( a = 2 \) and \( b = -2.5 \), the results from the cross-entropy algorithm are not consistent. Sometimes, it yields great results and the two final tilting parameters \((\alpha_1, \alpha_2)\) are where we expect them to be (B). Sometimes, the two final tilting parameters \((\alpha_1, \alpha_2)\) are almost indistinguishable. When this happens, simulation results are either seemingly “very accurate” with a very small variance and a large variance reduction (C), or “very inaccurate” with a variance much larger than that of the plain Monte Carlo estimate (D). This phenomenon has been observed in the literature [4, 8]. Indeed, the estimate in C looks very accurate with the 95% confidence interval being \(0.0227 \pm 0.0001\), which significantly underestimate the true value because it does not contain any (rare) samples that fall into the region \((-\infty, b)\). When
such rare samples do emerge, the variance rises drastically, since each such sample carries a huge likelihood ratio \((D)\). Similar simulation results are repeated in \(E\) and \(F\), except that in this case it does not even have a good run like \(B\).

In summary, the assignment of initial parameters by perturbation should be used with extreme caution, especially when different components of the mixture tend to collapse.

**Remark 2.** Table 1 also serves as a motivation for developing mixtures for importance sampling, since the simulation runs \(C, D, E,\) and \(F\) are essentially classical cross-entropy without using mixtures. One could easily be deceived by the “accurate” numerical results \((C\) and \(E\)).

**Remark 3.** It should be mentioned that when the estimation problem is associated with rare events, the naive initialization by perturbation will not work. For example, consider the case where the quantity of interest is a rare event probability \(P\{X \in A\}\), which corresponds to \(V(x) = 1_{\{x \in A\}}\). If one uses a small random perturbation to set up the initial parameters, then the denominators in the updating rule \((3.4)\) and \((3.2)\) will most likely become zero, rendering the first iteration and hence the scheme meaningless. Therefore, in this case, other methods of initialization, such as those discussed in the next two subsections, are needed.

### 4.2. Initialization by cross-entropy

An observation from the perturbation method is that if the initial assignment of the parameters is not too far away from optimality, the cross-entropy will most likely succeed. This leads to an approach that is very similar to [18, Chapter 8]. Even though their original algorithm is developed for estimating rare event probabilities, it can be modified to deal with mixtures.

The idea is to introduce a separate cross-entropy iterative scheme to gradually push the parameters towards reasonably good values. Note that the terminal value of the parameters from this separate cross-entropy scheme will become the initial assignment for the cross-entropy scheme in Section 3.

To describe this scheme, we observe that in many financial applications the function \(V(x)\) is of the form

\[
V(x) = H(x)1_{\{F(x) \in \bigcup_{j=1}^m A_j\}},
\]

where \(H(x)\) is nonnegative, \(F(x)\) is some function, and \(A_j\)'s are often associated with strike prices or barriers. Moreover, we assume that \(V\) can be embedded into a collection of functions (abusing notation) indexed by \(\delta = (\delta_1, \ldots, \delta_m)\):

\[
V_{\delta}(x) = H_{\delta}(x)1_{\{F(x) \in \bigcup_{j=1}^m A_j(\delta_j)\}}
\]

where \(H_{\delta}(x)\) is nonnegative and \(A_j(\delta_j)\) is a sequence of decreasing sets (with respect to \(\delta_j\)) for each \(j\) — for this reason, \(\delta\) is referred to as the rarity parameter. Without loss of generality, let \(V(x) = V_{\delta}(x)\) when \(\delta = 1 \triangleq (1, 1, \ldots, 1)\).

In order to avoid the collapse of the mixture, consider an increasing sequence

\[
\delta^1 \leq \delta^2 \leq \cdots \leq 1.
\]

If \(\delta^1\) is small, then with an initial parameter assignment from random perturbation, say \(\theta^i\), the cross-entropy scheme can identify a good mixture, say \(h_{\theta^1}\), for estimating

\[
E[V_{\delta^1}(X)].
\]
Using $\theta^i$ as the initial input, the cross-entropy scheme can then identify a good mixture, say $h_{\psi^2}$, for estimating

$$E[V_{\psi^2}(X)],$$

provided $\delta^2$ is not too far from $\delta^1$. The iterative scheme repeats these steps until a good mixture for

$$E[V_1(X)] = E[V(X)]$$
is found.

The choice of the sequence $\{\delta^i\}$ is to ensure that the mixture does not collapse, or that samples drawn from the mixture $h_{\psi^i}$ reach the set $A_j(\delta^{i+1})$ with non-trivial probability for each $j = 1, \ldots, m$. That is, we require

$$P\{F(X) \in A_j(\delta^{i+1})\} \geq \rho, \quad j = 1, \ldots, m,$$

where $X$ has the distribution $h_{\psi^i}$ and $0 < \rho < 1$ is some prefixed fraction. In this case, if the pilot sample size is $N$, then the number of samples to reach $A_j(\delta^{i+1})$ is approximately bounded from below

$$(4.3) \quad N \rho \cdot w,$$

where $w$ is the corresponding weight (it is a component of $\theta^i$).

Even though there are no explicit formula to determine $\{\delta^i\}$, they can be estimated by the same pilot samples that are used in the cross-entropy iterations. First of all, in all the iterations, the weights will be fixed at $w_1 = \ldots = w_m = 1/m$. The reason for that is (4.3) indicates the number of samples might be too few if the weights are allowed to adapt and decrease; see Remark 4. Secondly, suppose that $X_1, \ldots, X_N$ are pilot samples from $h_{\psi^i}$, then $\delta^{i+1}$ can be approximated by the largest $\delta$ such that the number of samples $\{F(X_1), \ldots, F(X_N)\}$ that belong to $A_j(\delta)$ is at least $N_0$ for each $j$, where $N_0 = \lceil N \rho / m \rceil$. Lastly, once $\delta^{i+1} = (\delta_1^{i+1}, \ldots, \delta_m^{i+1})$ is determined, $\theta^{i+1}$ can be obtained from equation (3.3) with $V(X_k)$ replaced by $V_{\delta^{i+1}}(X_k)$. The iteration will end if $\delta^{i+1} \geq 1$. More precisely, we have the following pseudocode.

**Pseudocode for the initialization of cross-entropy scheme:**

1. Choose a fraction $\rho \in (0, 1)$ and set $N_0 = \lceil N \rho / m \rceil$.
2. Set $\theta^0 = (1/m, \ldots, 1/m; \tilde{\alpha}_1^0, \ldots, \tilde{\alpha}_m^0)$ from some random perturbation.
3. Set iteration counter $i = 0$.
4. (*) Generate $N$ iid pilot samples $X_1, \ldots, X_N$ from density $h_{\psi^i}(x)$.
5. Set for each $j = 1, \ldots, m$,

$$\tilde{\delta}^{i+1} = \sup \{\delta : \#\{F(X_k) \in A_j(\delta)\} \geq N_0\} \vee \delta_j^i$$

6. Set $\delta^{i+1} = (\tilde{\delta}_1^{i+1}, \ldots, \tilde{\delta}_m^{i+1})$
7. For each $j = 1, \ldots, m$, set $\tilde{\alpha}_j^{i+1}$ as the solution to the equation

$$0 = \frac{1}{N} \sum_{k=1}^N V_{\delta^{i+1}}(X_k) f_{\delta^i}(X_k) \frac{\partial}{\partial \alpha} \log f_\alpha(X_k)$$

8. Set $\tilde{\theta}^{i+1} = (1/m, \ldots, 1/m; \tilde{\alpha}_1^{i+1}, \ldots, \tilde{\alpha}_m^{i+1})$
9. Set iteration counter $i = i + 1$.
10. If $\delta^{i+1} \geq 1$ then set $\theta^{i+1} = \theta^0$ and stop, otherwise go to (*).
Note that in the updating equation for $\delta$ we enforce that its components be nondecreasing as sometimes they can be driven down by the random noise, which might slow down the scheme. The parameter $\rho$ is usually chosen as a small number, but not too small so that $N_0$ is of medium size (a few hundreds usually suffice).

For illustration, consider the previous numerical example of estimating the probability in (4.1). One can let $\delta = (\delta_1, \delta_2)$ and

$$H_\delta(x) = 1, \quad A_1(\delta_1) = [\delta_1 a, \infty), \quad A_2(\delta_2) = (-\infty, \delta_2 b], \quad F(x) = x.$$ 

In this case, it is not difficult to check that the solution to the updating equation (4.4) for $\delta$ is

$$\bar{\delta}_1^{i+1} = \frac{X(N-N_0+1)}{a} \lor \bar{\delta}_1^i, \quad \bar{\delta}_2^{i+1} = \frac{X(N_0)}{b} \lor \bar{\delta}_2^i,$$

where $X(1) \leq X(2) \leq \cdots \leq X(N)$ is the order statistics of samples $X_1, \ldots, X_N$. Below are the numerical results. The initial parameter is set as before, i.e.,

$$\bar{\theta}^0 = (0.5, 0.5; 0, -0.1).$$

We also let $\rho = 0.05$. The entry $(\hat{a}_1^0, \hat{a}_2^0)$ in Table 2 is the terminal values of the tilting parameters from the initialization scheme. Note that the weights are always $(0.5, 0.5)$ in this initialization scheme, and

$$\hat{\theta}_0 = (0.5, 0.5; \hat{a}_1^0, \hat{a}_2^0)$$

will be the initial parameter setup for the iterative cross-entropy scheme. The entry “IT” records the number of iterations performed by the initialization scheme. We can see that the cross-entropy algorithm consistently yields accurate results with significant variance reduction.

| \{a, b\} | \{\hat{a}_1^0, \hat{a}_2^0\} | IT | CE Est | Var Ratio | \{w_1, w_2\} | \{a_1, a_2\} |
|----------|-----------------|----|---------|----------|-------------|-----------|
| \{(1, 1.5)\} | (1.13, 0.86) | 1 | 0.2255 | 2.4 | (0.69, 0.31) | (1.51, -1.87) |
| \{(2, 2.5)\} | (2.62, -3.12) | 4 | 0.0290 | 14.8 | (0.78, 0.22) | (2.38, -2.88) |
| \{(2, -3)\} | (2.69, -3.7) | 6 | 0.0240 | 17.1 | (0.95, 0.05) | (2.37, -3.29) |

Table 2. Estimating $P(X \geq a$ or $X \leq b)$ with CE initialization

The advantage of this initialization method is that it is rather mechanical and involves little analysis of the system. However, it requires some extra overhead on the computational budget. Moreover, numerical experiments seem to suggest that the terminal values of the tilting parameters from this initialization scheme sometimes are far from optimality, which slows down the convergence of the subsequent cross-entropy scheme.

Remark 4. The weights do not have to be fixed. It is sometimes more efficient to allow the weights to adapt as long as the minimal weight remains above some prespecified threshold.

4.3. Initialization by approximation. Yet another way to set up the initial mixture distribution $\hat{\theta}^0$ is by approximation, which is viable in many applications. For example, in many option pricing or risk management problems the function $V(X)$ is often nonzero only on sets of the from $\{R \geq a\}$ or union of such sets. The random variable $R$ can be functions of the underlying asset price, and the threshold $a$ can be some given strike price, barrier price, or loss threshold. The difficulty arises when $a$ is large. In this case, a rule of thumb is to choose a tilting parameter in an “economical”
way so that the original distribution is tilted just enough to ensure that \( \{ R \geq a \} \) is barely reached with nontrivial probability. This usually means to find an \( \alpha \)

\[
E_{f_{\alpha}}[R] = a.
\]

If we do this for each of such sets, then we have an initial assignment of \((\hat{\alpha}_1^0, \ldots, \hat{\alpha}_m^0)\).

Again, we use problem (4.1) to illustrate this approach. In this case, what the initialization amounts to is to shift the mean of the new distribution to \(a\) or \(b\). That is,

\[
\hat{\theta}^0 = (0.5, 0.5; a, b).
\]

The results are similar to that of Table 2.

| \(\{a, b\}\) | CE Est | Var Ratio | \((w_1, w_2)\) | \((\alpha_1, \alpha_2)\) |
|----------|--------|-----------|----------------|-----------------|
| \(-1.5, 1\) | 0.2258 | 2.3 | (0.70, 0.30) | (1.51, -1.92) |
| \(-2.5, 2\) | 0.0290 | 14.0 | (0.78, 0.22) | (2.37, -2.82) |
| \(-3, 3\) | 0.0241 | 17.2 | (0.86, 0.05) | (2.37, -3.27) |

Table 3. Estimating \(P\{X \geq a \) or \(X \leq b\)\)

The initialization by approximation requires some analysis of the system, which may be hard to carry out at times. However, if such an approximation is available, the convergence of the corresponding iterative cross-entropy is usually very fast.

5. Numerical Examples. In this section we include a number of numerical experiments. In all the simulations, the pilot sample size is \(N = 10000\), the sample size for importance sampling is \(n = 100000\), and number of iterations in the cross-entropy algorithm is set to be \(IT_{\text{NUM}} = 5\), unless otherwise specified. In all the tables, “CE Est” stands for the estimate from importance sampling using the cross-entropy method, “R.E.” denotes the empirical relative error, which is defined to be

\[
\text{empirical relative error} = \frac{\text{standard error of estimate}}{\text{estimate}}.
\]

and “Var Ratio” is again the ratio of the variance of the plain Monte Carlo estimate to that of the cross-entropy estimate.

EXAMPLE 1. We consider a very simple example of estimating the price of an average price call option. More precisely, let the stock price be

\[
\frac{dS_t}{S_t} = r \, dt + \sigma \, dW_t
\]

under the risk-neutral probability measure, where \(r\) is the risk-free interest rate and \(W\) a standard Brownian motion. Consider a discretely monitored average price call option with payoff \((\bar{S} - K)^+\) and maturity \(T\), where \(\bar{S}\) is the arithmetic mean

\[
\bar{S} = \frac{1}{d} \sum_{i=1}^{d} S_{t_i}
\]

for a given set of dates \(0 < t_1 < \cdots < t_d = T\).

Since the option payoff is positive on a simple set \(\{ \bar{S} \geq K \} \), we will use importance sampling with the classical cross-entropy method, that is, without mixtures. Define for \(1 \leq i \leq d\),

\[
Z_i = \frac{(W_{t_i} - W_{t_{i-1}})}{\sqrt{t_i - t_{i-1}}}
\]

Then \(X = (Z_1, \ldots, Z_d)\) is a \(d\)-dimensional
standard normal random vector. The stock price at time $t_i$ can be written as a function of $X$:

$$S_{t_i} = S_0 \exp \left\{ \left( r - \frac{1}{2} \sigma^2 \right) t_i + \sigma \sum_{j=1}^{i} \sqrt{t_j - t_{j-1}} Z_j \right\}.$$ 

Therefore, the discounted option payoff $e^{-rT}(\bar{S} - K)^+$ is a function of $X$ as well, say $V(X)$, and the option price is $E[V(X)]$.

Since the updating rule, or the solution to equation (2.6), is explicitly given by Lemma 2.1, the only remaining question is the initial parameter $\hat{\theta}_0 = (\hat{\theta}_0^1, \ldots, \hat{\theta}_0^d) \in \mathbb{R}^d$. If the strike price $K$ is not large, one can just let $\hat{\theta}_0 = (0, \ldots, 0)$, that is, starting with the original distribution. However, if $K$ increases, this initialization will not be adequate because fewer and fewer sample path will reach the strike price $K$; see Remark 3. In this case, we will need a different initialization. One may use the general initialization method by cross-entropy, which is sufficient and rather straightforward. Here we illustrate how to use the method of approximation. What we would like is that $\bar{S}$ will exceed $K$ with nontrivial probability if $X = (Z_1, \ldots, Z_d)$ is distributed as $N(\hat{\theta}_0^0, I_d)$. This leads to equation

$$K = E[\bar{S}] = \frac{1}{d} \sum_{i=1}^{d} S_0 \exp \left\{ \left( r - \frac{1}{2} \sigma^2 \right) t_i + \sigma \sum_{j=1}^{i} \sqrt{t_j - t_{j-1}} \hat{\theta}_0^j \right\}.$$ 

A convenient choice is to let $\hat{\theta}_0 = (a, a, \ldots, a)$, where $a$ is the solution to the equation

$$(5.1) \quad K = \frac{1}{d} \sum_{i=1}^{d} S_0 \exp \left\{ \left( r - \frac{1}{2} \sigma^2 \right) t_i + \sigma \sum_{j=1}^{i} \sqrt{t_j - t_{j-1}} \cdot a \right\}.$$ 

It can be easily solved by numerical methods.

Some numerical results are presented in Table 4. The model parameters are given by

$$S_0 = 50, \quad r = 0.05, \quad \sigma = 0.3, \quad T = 1, \quad d = 30, \quad t_i = \frac{i}{d}.$$ 

The initial tilting parameter are obtained by solving equation (5.1) with the bisection method.

| Strike | CE Est | CE Est | CE Est | CE Est | CE Est |
|--------|--------|--------|--------|--------|--------|
| $K = 50$ | 4.0766 | 4.0766 | 4.0766 | 4.0766 | 4.0766 |
| $K = 60$ | 1.0179 | 1.0179 | 1.0179 | 1.0179 | 1.0179 |
| $K = 70$ | 0.1917 | 0.1917 | 0.1917 | 0.1917 | 0.1917 |
| $K = 80$ | 0.0309 | 0.0309 | 0.0309 | 0.0309 | 0.0309 |
| $K = 90$ | 0.0045 | 0.0045 | 0.0045 | 0.0045 | 0.0045 |

Table 4. Average price call option

Clearly the variance reduction is more significant when the option becomes more out-of-the-money. This is a common phenomenon in importance sampling, which is most efficient in the dealing with rare events.

**Example 2.** Consider a rainbow option with $d$-underlying assets, whose prices are modeled by geometric Brownian motions under the risk-neutral probability measure:

$$\frac{dS_{t}^{(j)}}{S_{t}^{(j)}} = rt + \sigma_j dW_t^{(j)}, \quad j = 1, \ldots, d.$$
Here $r$ is the risk-free interest rate, and $W = (W^{(1)}, \ldots, W^{(d)})$ is a $d$-dimensional Brownian motion with covariance matrix $\Sigma = [\Sigma_{ij}]$ such that $\Sigma_{jj} = 1$ for $j = 1, \ldots, d$. We are interested in estimating the prices of an outperformance option with maturity $T$ and payoff

$$\left( \max \left\{ S_T^{(1)}, \ldots, S_T^{(d)} \right\} - K \right)^+. $$

Let $C$ be a Cholesky factorization of $\Sigma$, that is, $C$ is a lower triangular matrix such that $CC' = \Sigma$. We can write

$$\frac{1}{\sqrt{T}} W_T = \frac{1}{\sqrt{T}} (W_T^{(1)}, \ldots, W_T^{(d)}) = CX$$

for some $d$-dimensional standard normal random vector $X$. The discounted option payoff is just $V(X)$, where

$$V(x) = \left( \max_{j=1,\ldots,d} S_0^{(j)} \exp \left\{ -\frac{1}{2} \sigma_j^2 T + \sigma_j \sqrt{T} (Cx)_j \right\} - e^{-rT} K \right)^+$$

for $x \in \mathbb{R}^d$.

Since the option has strictly positive payoff when either of the $d$ stock prices at maturity exceeds the strike price $K$, it is natural to use a mixture with $m = d$ components. The updating rule is explicitly given by equations (3.4)–(3.5) and Lemma 3.1. Therefore, the only question remaining is the initialization. We will illustrate both methods, i.e., initialization by cross-entropy and by approximation.

(i) **Initialization by cross-entropy:** Rewrite the discounted option payoff as

$$V(X) = H(X) 1_{\{F(X) \in \bigcup_{j=1}^{d} A_j\}}$$

where $H(x) = V(x)$ and $F(x) = x$ for $x \in \mathbb{R}^d$, and

$$A_j = \left\{ x \in \mathbb{R}^d : S_0^{(j)} \exp \left\{ \left( r - \frac{1}{2} \sigma_j^2 \right) T + \sigma_j \sqrt{T} (Cx)_j \right\} > K \right\}.$$ 

We can embed it into

$$V_\delta(X) = H_\delta(X) 1_{\{F(X) \in \bigcup_{j=1}^{d} A_j(\delta_j)\}}$$

where

$$H_\delta(x) = \left( \max_{j=1,\ldots,d} S_0^{(j)} \exp \left\{ -\frac{1}{2} \sigma_j^2 T + \sigma_j \sqrt{T} (Cx)_j \right\} - e^{-rT} \delta_j K \right)^+$$

and $A_j(\delta_j)$ is defined exactly as $A_j$ except that $K$ is replaced by $\delta_j K$. The pseudocode for the initialization is given in Section 4.2 where the updating equation (4.4) for $\delta$ can be explicitly solved as follows: Let $X_1, \ldots, X_N$ be the pilot samples in the $i$-th iteration. For each $j = 1, \ldots, d$, define a sequence (which is indeed the corresponding stock prices)

$$Y_k^{(j)} = S_0^{(j)} \exp \left\{ \left( r - \frac{1}{2} \sigma_j^2 \right) T + \sigma_j \sqrt{T} (CX_k)_j \right\}, \quad k = 1, \ldots, N.$$
Denote its order statistics by $Y_{(1)} \leq \cdots \leq Y_{(N)}$. It is not difficult to see that the solution to the updating equation (4.4) is

$$\tilde{\delta}_{j}^{i+1} = \frac{1}{K} Y_{(j)}^{(i)} \vee \tilde{\delta}_{j}^{i}.$$ 

(ii) **Initialization by approximation:** The initial assignment of the tilting parameters $(\hat{\alpha}_{1}^{0}, \ldots, \hat{\alpha}_{d}^{0})$ is such that

$$E_{\hat{\alpha}_{0}^{0}}[S_{T}^{(j)}] = K.$$ 

That is, if $X$ is distributed as $N(\hat{\alpha}_{j}^{0}, \Sigma)$, then

$$E[S_{0}^{(j)} \exp \left\{ (r - \frac{1}{2} \sigma_{j}^{2}) T + \sigma_{j} \sqrt{T} (C X)_{j} \right\}] = K,$$

or equivalently,

$$S_{0}^{(j)} \exp \left\{ r T + \sigma_{j} \sqrt{T} (C \hat{\alpha}_{j}^{0})_{j} \right\} = K.$$ 

There are many solutions to this equation. It is convenient to let

$$\hat{\alpha}_{j}^{0} = C^{-1} \eta_{j},$$

where $\eta_{j} \in \mathbb{R}^{d}$ and all its components are zero except the $j$-th component which equals

$$\frac{1}{\sigma_{j} \sqrt{T}} \left( \log \frac{K}{S_{0}^{(j)}} - r T \right).$$

In the numerical experimentation, we include results from both methods of initialization. The entry “INI CE” denotes the results using initialization by cross-entropy, while “INI AP” denotes those via initialization by approximation.

The first numerical example is concerned with a two-stock outperformance option with maturity $T = 1$ and $S_{0}^{(1)} = 50$, $S_{0}^{(2)} = 45$, $r = 0.03$, $\sigma_{1} = 0.1$, $\sigma_{2} = 0.15$, $\Sigma = \begin{bmatrix} 1 & 0.2 \\ 0.2 & 1 \end{bmatrix}$.

The numerical results are reported in Table 5. The two initialization methods yield almost identical estimates and relative errors.

| Strike | $K = 50$ | $K = 60$ | $K = 70$ |
|--------|-----------|-----------|-----------|
| INI CE | INI AP    | INI CE    | INI AP    | INI CE    | INI AP    |
| CE Est | 3.5898    | 3.5825    | 0.2768    | 0.2763    | 0.0093    | 0.0093    |
| R.E.   | 0.14%     | 0.14%     | 0.21%     | 0.28%     | 0.37%     | 0.37%     |
| Var Ratio | 6.2%   | 5.2%      | 27.2      | 26.9      | 445.8     | 456.5     |

Table 5: Outperformance option with two underlying assets

The second numerical example has $d = 4$ underlying assets with the following parameters.

$S_{0}^{(1)} = 45$, $S_{0}^{(2)} = 50$, $S_{0}^{(3)} = 47$, $S_{0}^{(4)} = 50$, $r = 0.02$, $T = 0.5$. 


\( \sigma_1 = \sigma_2 = 0.1, \sigma_3 = \sigma_4 = 0.2, \quad \Sigma = \begin{bmatrix} 1.0 & 0.3 & -0.2 & 0.4 \\ 0.3 & 1.0 & -0.3 & 0.1 \\ -0.2 & -0.3 & 1.0 & 0.5 \\ 0.4 & 0.1 & 0.5 & 1.0 \end{bmatrix} \). We also let IT\_NUM = 10 in the simulation because of the slow convergence when initialization by cross-entropy is adopted. The numerical results are reported in Table 6.

| Strike | \( K = 50 \) | \( K = 60 \) | \( K = 70 \) |
|--------|---------------|---------------|---------------|
| CE Est | 4.6841        | 4.6722        | 0.5271        |
| R.E.   | 0.17\%        | 0.17\%        | 0.26\%        |
| Var Ratio | 3.5           | 3.5           | 19.9          |

Table 6: Outperformance option with four underlying assets

Even though both methods lead to significant variance reduction, more extensive numerical experiments appear to indicate that the convergence of the iterative cross-entropy algorithm is much faster when combined with the method of initialization by approximation. This is because initialization by approximation usually gives a set of parameters closer to the optimal one.

**Example 3.** The setup is the same as the previous example. We are interested in estimating the price of a pyramid rainbow option with maturity \( T \) and payoff

\[
\left( |S_T^{(1)} - K_1| + \cdots + |S_T^{(d)} - K_d| - K \right)^+ \]

where \( K_1, \ldots, K_d \) are all positive constants.

It is not difficult to see that the payoff of the option is positive on each of the following \( 2^d \) sets:

\[
\left\{ \pm |S_T^{(1)} - K_1| \geq 0, \ldots, \pm |S_T^{(d)} - K_d| \geq 0 : \sum_{j=1}^d \pm |S_T^{(j)} - K_j| \geq K \right\}.
\]

In other words, in each set and for each \( j = 1, \ldots, d \), it is either \( S_T^{(j)} \geq K_j \) or \( S_T^{(j)} \leq K_j \), and the absolute values are expressed accordingly. Therefore, it is natural to select \( m = 2^d \). As for initialization, one can use the method by cross-entropy as before, but the convergence is slow. Instead, we will use the method by approximation, which is fairly easy to carry out and the convergence is much faster. To this end, recall the Cholesky factorization \([5.2]\) and rewrite

\[
S_T^{(j)} = S_0^{(j)} \exp \left\{ \left( r - \frac{1}{2} \sigma_j^2 \right) T + \sigma_j \sqrt{T} (CX)_j \right\},
\]

where \( X \) is a standard \( d \)-dimensional normal random vector. Consider a typical set, say

\[
\left\{ S_T^{(1)} \geq K_1, \ldots, S_T^{(d)} \geq K_d : \sum_{j=1}^d [S_T^{(j)} - K_j] \geq K \right\}.
\]

For this set, we would like to construct a tilting parameter \( \hat{\alpha} \in \mathbb{R}^d \) such that if \( X \) is distributed as \( N(\hat{\alpha}, I_d) \), then the above set is not a rare event. Note that under the
original distribution,
\[ E[S_T^{(j)}] = S_0^{(j)} \exp\{rT\}, \]
while under the new distribution
\[ E_\hat{\alpha} [S_T^{(j)}] = S_0^{(j)} \exp \left\{ rT + \sigma_j \sqrt{T} (C \hat{\alpha})_j \right\} := x_j. \]

The idea is to push \( x_j \) beyond \( K_j \) if \( E[S_T^{(j)}] < K_j \), which means we would require
\[ x_j \geq \max \left\{ K_j, S_0^{(j)} \exp\{rT\} \right\} := \bar{K}_j. \]
Obviously we should also require
\[ \sum_{j=1}^{d} (x_j - K_j) \geq K. \]

A simple choice is to let \( x_j = \bar{K}_j + a \) for some constant \( a \geq 0 \). Then (5.4) becomes
\[ a \geq \frac{1}{d} \left[ K - \sum_{j=1}^{d} (\bar{K}_j - K_j) \right]. \]
The formula of \( x_j \) for a general set can be obtained in exactly the same way (we omit the details) and is indeed
\[ x_j = \bar{K}_j + \frac{1}{d} \left[ K - \sum_{i=1}^{d} (\bar{K}_i - K_i) \right]^+. \]

Below are some numerical results. The first example has \( d = 2 \) underlying assets. Thus we use mixtures with \( m = 2^d = 4 \) components. The parameters are
\[ S_0^{(1)} = 50, \quad S_0^{(2)} = 45, \quad K_1 = 55, \quad K_2 = 50, \quad r = 0.03, \quad T = 1, \]
\[ \sigma_1 = 0.2, \quad \sigma_2 = 0.25, \quad \Sigma = \begin{bmatrix} 1 & 0.3 \\ 0.3 & 1 \end{bmatrix}. \]
The simulation results are reported in Table 7.
The second example considers a model with \( d = 4 \) underlying assets. The number of components in the mixture is \( m = 2^d = 16 \). The parameters are

\[
S_0^{(1)} = 50, \quad S_0^{(2)} = 45, \quad S_0^{(3)} = 45, \quad S_0^{(4)} = 30, \quad T = 1,
\]

\[
K_1 = 55, \quad K_2 = K_3 = 50, \quad K_4 = 35, \quad r = 0.03, \quad \sigma_1 = \sigma_2 = 0.15,
\]

\[
\sigma_3 = \sigma_4 = 0.2, \quad \Sigma = \begin{bmatrix} 1 & 0.1 & -0.2 & 0.3 \\ 0.1 & 1 & -0.5 & 0.4 \\ -0.2 & -0.5 & 1 & 0.2 \\ 0.3 & 0.4 & 0.2 & 1 \end{bmatrix}.
\]

The simulation results are reported in Table 8.

### Example 4.
Consider a two dimensional CEV model where the stock prices satisfy the following stochastic differential equations under the risk-neutral probability measure:

\[
dS_t = rS_t \, dt + \sigma_1 S_t^{\gamma_1} \, dW_t,
\]

\[
dH_t = rH_t \, dt + \sigma_2 H_t^{\gamma_2} \, dB_t,
\]

where \( \gamma_1, \gamma_2 \in [0.5, 1) \) are given constants, and \((W, B)\) is a two-dimensional Brownian motion with covariance matrix

\[
\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.
\]

We are interested in estimating the price of a digital call option with payoff

\[
1_{\{\max\{c_1 S_t, c_2 H_t\} \geq K}\}}
\]

with maturity \( T \).

We work with the discounted stock prices, which makes the future analysis slightly easier. More precisely, let \( X_t = e^{-rt} S_t \) and \( Y_t = e^{-rt} H_t \). It follows from Itô formula that

\[
\begin{align*}
\frac{dX_t}{X_t} &= \sigma_1 e^{-r(1-\gamma_1)t} X_t^{\gamma_1} \, dW_t, \\
\frac{dY_t}{Y_t} &= \sigma_2 e^{-r(1-\gamma_2)t} Y_t^{\gamma_2} \, dB_t.
\end{align*}
\]

Since there are two ways to have a positive payoff, i.e., either of the two stock prices reaching a high enough level at maturity, we let \( m = 2 \). The question is how to...
determine the initial tilting parameters. One can use initialization by cross-entropy, but the convergence is very slow. So here we use the method of approximation again.

Consider a simple change of measure $Q_1$ under which $W$ becomes a Brownian motion with constant drift $x$. The drift $x$ will be determined so that the equation

$$E_{Q_1}[X_T] = e^{-rT}E_{Q_1}[S_T] = e^{-rT}K/c_1$$

is satisfied approximately, which ensures $c_1S_T$ to reach the target level $K$ with non-trivial probability. To this end, we rewrite equation (5.5) as

$$dX_t = x\sigma_1 e^{-r(1-\gamma_1)t}X_t^{\gamma_1} dt + \sigma_1 e^{-r(1-\gamma_1)t}X_t^{\gamma_1} d(W_t - xt),$$

where $W_t - xt$ is a Brownian motion under probability measure $Q_1$. Taking expected values on both sides and using Hölder inequality, we arrive at

$$E_{Q_1}[X_t] = X_0 + \int_0^t x\sigma_1 e^{-r(1-\gamma_1)s}E[X_s^{\gamma_1}] ds$$

$$\geq X_0 + \int_0^t x\sigma_1 e^{-r(1-\gamma_1)s} (E[X_s])^{\gamma_1} ds.$$ 

This motivates the approximation of $E_{Q_1}[X_t]$ by the function $f(t)$, which is defined to be the solution to the equation

$$(5.7) \quad f(t) = X_0 + \int_0^t x\sigma_1 e^{-r(1-\gamma_1)s} f^{\gamma_1}(s) ds.$$ 

It can be shown that $f(t)$ serves as a lower bound for $E_{Q_1}[X_t]$. Actually, it is also a very good approximation for the following reason. Note that $E[X_s^{\gamma_1}] = (E[X_s])^{\gamma_1}$ when $\gamma_1 = 1$ or when $X_s$ is a constant (i.e., variance of $X_s$ is zero). Therefore, when $\gamma_1$ is close to one, we expect $f(s)$ to approximate $E_{Q_1}[X_s]$ well. When $\gamma_1$ is away from one, the variance of $X_s$ is often relatively small, which again implies that $f(s)$ is a good approximation. It is natural now to choose $x$ so that $f(T) = e^{-rT}K/c_1$. To this end, observe that equation (5.7) is equivalent to the ordinary differential equation

$$\frac{1}{f^{\gamma_1}} \frac{df}{dt} = x\sigma_1 e^{-r(1-\gamma_1)t}, \quad f(0) = X_0 = S_0.$$ 

Solving this equation explicitly we obtain

$$f^{1-\gamma_1}(t) - S_0^{1-\gamma_1} = \frac{x\sigma_1}{r} \left[ 1 - e^{-r(1-\gamma_1)t} \right].$$

Letting $f(T) = e^{-rT}K/c_1$, we arrive at

$$(5.8) \quad x = \frac{r \left( e^{-rT}K/c_1 \right)^{1-\gamma_1} - S_0^{1-\gamma_1}}{1 - e^{-r(1-\gamma_1)T}}.$$ 

Note that $dB_t = \rho dW_t + \sqrt{1 - \rho^2} dB_t$ where $W$ and $B$ are two independent standard Brownian motions. The probability measure $Q_1$ shifts the drift of $W$ to $x$, but $B_t$ will remain a standard Brownian motion under $Q_1$. 

We use the following version of Euler scheme to discretize the processes under $Q_1$. Given a large integer $N_0$, let $\Delta t = T/N_0$ and $t_i = i\Delta t$ for $i = 0, \ldots, N_0$. Let
\[ X_0 = X_0 = S_0, Y_0 = Y_0 = H_0, \] and recursively define

\[
\begin{align*}
\dot{X}_{t+1} &= \dot{X}_{t_i} + \sigma_1 e^{-r(1-\gamma_1)t_i} \dot{X}_{t_i} \gamma_1 \Delta t + \sigma_1 e^{-r(1-\gamma_1)t_i} \dot{X}_{t_i} \gamma_1 \sqrt{\Delta t} \cdot Z_{t+1}, \\
\dot{Y}_{t+1} &= \dot{Y}_{t_i} + \rho \sigma_2 e^{-r(1-\gamma_2)t_i} \dot{Y}_{t_i} \gamma_2 \Delta t + \rho \sigma_2 e^{-r(1-\gamma_2)t_i} \dot{Y}_{t_i} \gamma_2 \sqrt{\Delta t} \cdot Z_{t+1} + \sigma_2 e^{-r(1-\gamma_2)t_i} \dot{Y}_{t_i} \sqrt{\Delta t} \cdot \sqrt{1 - \rho^2 R_{t+1}},
\end{align*}
\]

where \{Z_t, R_t\} are iid standard normal random variables. In the implementation, if \(\dot{X}_{t_i}\) or \(\dot{Y}_{t_i}\) becomes negative, we reset it to zero.

The second change of measure \(Q_2\) can be determined similarly. Under \(Q_2\), \(B\) becomes a Brownian motion with drift \(y\), which, analogous to (5.8), is given by

\[
y = \frac{r}{\sigma_2} \left( e^{-rT} K/c_2 \right)^{1-\gamma_2} - S_0^{1-\gamma_2},
\]

and the corresponding discretization scheme is

\[
\begin{align*}
\dot{X}_{t+1} &= \dot{X}_{t_i} + \rho y \sigma_1 e^{-r(1-\gamma_1)t_i} \dot{X}_{t_i} \gamma_1 \Delta t + \rho \sigma_1 e^{-r(1-\gamma_1)t_i} \dot{X}_{t_i} \gamma_1 \sqrt{\Delta t} \cdot Z_{t+1}, \\
\dot{Y}_{t+1} &= \dot{Y}_{t_i} + y \sigma_2 e^{-r(1-\gamma_2)t_i} \dot{Y}_{t_i} \gamma_2 \Delta t + \sigma_2 e^{-r(1-\gamma_2)t_i} \dot{Y}_{t_i} \gamma_2 \sqrt{\Delta t} \cdot Z_{t+1},
\end{align*}
\]

where \{Z_t, R_t\} are iid standard normal random variables. Again, if \(\dot{X}_{t_i}\) or \(\dot{Y}_{t_i}\) becomes negative, we reset it to zero.

Some numerical results are reported in Table 9. The parameters are given by

\[
S_0 = 50, \ H_0 = 48, \ \sigma_1 = 0.3, \ \sigma_2 = 0.35, \ \gamma_1 = 0.5, \ \gamma_2 = 0.7, \ \rho = 0.3,
\]

\[
c_1 = c_2 = 1, \ r = 0.03, \ T = 1, \ N_0 = 50.
\]

| Strike | \(K = 50\) | \(K = 55\) | \(K = 60\) | \(K = 65\) | \(K = 70\) |
|--------|------------|------------|------------|------------|------------|
| CE Est | 0.8297     | 0.1908     | 0.0314     | 0.0039     | 3.3638 × 10^{-4} |
| R.E.   | 0.14%      | 0.37%      | 0.48%      | 0.56%      | 0.63%      |
| Var Ratio | 1.0      | 3.0       | 13.8      | 82.4      | 545.5     |

Table 9. Digital options for a CEV model

REFERENCES

[1] S. Asmussen. Ruin Probabilities. World Scientific, Singapore, 2000.
[2] J. Blanchet and P. Glynn. Efficient rare-event simulation for the maximum of heavy-tailed random walks. Ann. Appl. Prob., 18(2008), pp. 1351–1378.
[3] J. Blanchet and Y. Shi. Strongly efficient cross-entropy method for heavy-tailed simulation. Proceedings of the 2011 Winter Simulation Conference, S. Jain, R.R. Creasey, J. Himmelspach, K.P. White, and M. Fu, eds., 2011.
[4] T.I. Brereton, J.C.C. Chan, and D.P. Kroese. Fitting mixture importance sampling distributions via improved cross-entropy. Proceedings of the 2011 Winter Simulation Conference, S. Jain, R.R. Creasey, J. Himmelspach, K.P. White, and M. Fu, eds., 2011.
[5] J.A. Bucklew. Introduction to Rare Event Simulation. Springer-Verlag, New York, 2010.
[6] P.T. De Boer. Analysis of state-independent importance sampling measures for the two-node tandem queue. ACM Trans. Modeling Comp. Simulation, 16 (2006), pp. 225–250.
[7] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society Series B, 39 (1977), pp. 1–38.
[8] P. Dupuis and H. Wang. Importance sampling, large deviations, and differential games. Stoch. and Stoch. Reports., 76(2004), pp. 481–508.
[9] P. Dupuis and H. Wang. Subsolutions of an Isaacs equation and efficient schemes for importance sampling. Math. Oper. Res., 32(2007), pp. 1–35.
[10] P. Dupuis and H. Wang. Importance sampling for Jackson networks. Queueing Systems, 62(2009), pp. 113–157.
[11] P. Glasserman and Y. Wang. Counter examples in importance sampling for large deviations probabilities. Ann. Appl. Prob., 7(1997), pp. 731–746.
[12] P. Glasserman. Monte Carlo Methods in Financial Engineering. Springer-Verlag, New York, 2004.
[13] P. Glasserman and P. Heidelberger and P. Shahabuddin. Asymptotically optimal importance sampling and stratification for pricing path-dependent options. Mathematical Finance, 9(1999), pp. 117–152.
[14] P. Guasoni and S. Robertson. Optimal Importance Sampling with Explicit Formulas in Continuous Time. Finance and Stochastics, 12(2008), pp. 1–19.
[15] P. Heidelberger. Fast simulation of rare events in queueing and reliability models. ACM Trans. Modeling Comp. Simulation, 4(1995), pp. 43–85.
[16] R.Y. Rubinstein. Optimization of computer simulation models with rare events. Euro. J. Oper. Res, 99(1997), pp. 89–112.
[17] R.Y. Rubinstein. The simulated entropy method for combinatorial and continuous optimization. Methodology and Computing in Applied Probability, 2(1999), pp. 127–190.
[18] R.Y. Rubinstein and D.P. Kroese. Simulation and the Monte Carlo Method. John Wiley & Sons, Hoboken, New Jersey, 2007.
[19] J.S. Sadowsky. On Monte Carlo estimation of large deviations probabilities. Ann. Appl. Prob., 6(1996), pp. 399–422.
[20] J. Sadowsky and J.A. Bucklew. On large deviations theory and asymptotically efficient Monte Carlo estimation. IEEE transactions on Information Theory, 36(1990), pp. 579–588.
[21] P. Shahshuddin. Importance sampling for the simulation of highly reliable Markovian systems. Management Sciences, 40(1994), pp. 333–352.
[22] D. Siegmund. Importance sampling in the Monte Carlo study of sequential tests. Ann. Statist., 4(1976), pp. 673–684.
[23] H. Wang. Monte Carlo Simulation with Applications to Finance. CRC Press, Taylor & Francis Group, Florida, 2012.