University "Alexandru Ioan Cuza" of Iaşi
Faculty of Computer Science

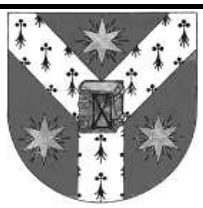
TECHNICAL REPORT

# Importance Sampling using Rényi divergence

**Emanuel Florentin Olariu**

**TR 12-02,** May 2012

**ISSN 1224-9327**



**Universitatea "Alexandru Ioan Cuza" Iaşi**
**Facultatea de Informatică**
Str. Berthelot 16, 6600-Iaşi, Romania
Tel. +40-32-201090, email: bibl@infoiasi.ro

**Abstract**

We present an alternative approach to the problem of estimating probabilities of rare events and for optimization problems using the class of Rényi divergences of order $\alpha > 1$. The general procedure we describe does not involve any specific family of distributions, the only restriction is that the search space consists of product form probability density functions. We discuss an algorithm for estimation of probability of rare events and a version for continuous optimization. The results of numerical experimentation with these algorithms carried in the last section support their performances.

**Keywords:** Rényi divergence, Monte Carlo, importance sampling, convex optimization.

# 1 Introduction

Many problems which arise in a variety of applications of operations research can be described as the evaluation of the expected value for a given random variable. Areas of interest which use such an evaluation are rare event simulation or global optimization. The problem discussed here is the estimation of

$$m = \mathbb{E}_f\left[\mathcal{H}(\mathbf{X})\right] = \int_\Omega \mathcal{H}(s) f(s)\, d\mu(s), \tag{1}$$

where $\mathbf{X} \in \mathbb{R}^n$ has a probability density function (pdf) $f$, and $\mathcal{H} : \mathbb{R}^n \to \mathbb{R}_+$ is a Lebesgue measurable (integrable) function. A known method for estimating $m$ is Monte Carlo basic simulation which gives an unbiased estimator

$$m_N = \frac{1}{N}\sum_{i=1}^{N} \mathcal{H}(X^i),$$

where $\left(X^i\right)_{1\leqslant i\leqslant N}$ are independent and identical distributed (i.i.d.) samples from $f$. There are many methods which involves variance minimization for the Monte Carlo estimator (see [5]). One of the most known is the importance sampling which chooses an alternate pdf, say $g$, such as $supp(f\cdot\mathcal{H}) \subseteq supp(g)$, and estimate $m$ by

$$m_N(g) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{H}(X^i)\frac{f(X^i)}{g(X^i)}, \tag{2}$$

where $\left(X^i\right)_{1\leqslant i\leqslant N}$ are independent and identical distributed (i.i.d.) samples from $g$. $m_N(g)$ is also an unbiased estimator of $m$, and $g$ is known as the importance sampling (IS) distribution. The optimal IS distribution, in order to achieve a zero-variance estimator, is

$$g^*(s) = \frac{\mathcal{H}(s)f(s)}{m} \tag{3}$$

This pdf is hard to determine as it depends on the desired value $m$; a more practical approach is to search for an IS distribution from a parametric family $(\overline{g}_\theta)_\theta$ (such the natural exponential family) which has to minimize the Kullback-Leibler divergence (see [2]). In many situations this search is refined and we look from distributions of the form $\overline{g}(s) = \overline{g}_1(s_1)\cdot\ldots\cdot\overline{g}_n(s_n)$ - see [2], [6], [7] and [9]. In this framework the chosen IS distribution $\overline{g}$ is a solution to the following problem

$$\min_{g\in\mathcal{G}} D(g^*||g) = \min_{g\in\mathcal{G}}\left(\int_{\mathbb{R}^n} g^*(s)\ln\frac{g^*(s)}{g(s)}\, d\mu(s)\right), \tag{4}$$

where

$$\mathcal{G} = \left\{ g : \mathbb{R}^n \to \mathbb{R}_+ \,:\, g(s) = \prod_{i=1}^{n} g_i(s_i), \forall s\in\mathbb{R}^n, \int_{\mathbb{R}} g_i(s_i)\, d\mu(s_i) = 1, \forall i \right\}$$

We investigate an alternative method to the cross entropy procedure, based on the *Rényi divergence of order* $\alpha$. Let $(\Omega, \mathcal{H}, \mu)$ a probability space, $p$ and $q$ two pdfs, and $\alpha \in \mathbb{R}_+^* \setminus \{1\}$, the Rényi divergence of $p$ and $q$ is

$$D_\alpha(p||q) = \frac{1}{\alpha-1}\ln\left(\int_\Omega \left[p^\alpha(s)q^{1-\alpha}(s)\right] d\mu(s)\right) \tag{5}$$

It is known from [4] that $\lim_{\alpha\uparrow 1} D_\alpha(p||q) = D(p||q)$, where $D(p||q)$ is the *Kullback-Leibler divergence*. In this sense the Rényi divergence is a generalization of those of Kullback and Leibler.

## 2 Minimizing the Rényi divergence

In the following sections we suppose that $\alpha > 1$. As mentioned earlier we propose to choose as IS distribution which minimize the Rényi divergence:

$$\min_{g\in\mathcal{G}}\left(\int_{\mathbb{R}^n}\left[g^{*\alpha}(s)g^{1-\alpha}(s)\right]\,d\mu(s)\right) = \min_{g\in\mathcal{G}}\left(\int_{\mathbb{R}^n}\left[\mathcal{H}^\alpha(s)f^\alpha(s)g^{1-\alpha}(s)\right]\,d\mu(s)\right) \tag{6}$$

For a given $\varepsilon_0 \in (0,1)$, say $\varepsilon_0 = 1/2$, let us denote by

$$U = \left\{h : \mathbb{R}^n \to \mathbb{R}_+ \,:\, h \in L^1(\mathbb{R})\right\}, U_0 = \left\{h \in U \,:\, \left|\int_{\mathbb{R}} h(t)\,d\mu(t) - 1\right| < \varepsilon_0\right\};$$

$U_0$ and $U_0^n$ are convex subsets of the Banach spaces $L^1(\mathbb{R})$ and $\left(L^1(\mathbb{R})\right)^n$, respectively. (6) becomes the following functional minimization problem

$$\min_{g\in U_0^n}\left(\int_{\mathbb{R}^n}\left[\mathcal{H}^\alpha(s)f^\alpha(s)\prod_{i=1}^n g_i^{1-\alpha}(s_i)\right]\,d\mu(s)\right),$$
$$\text{subject to } \int_{\mathbb{R}} g_i(s_i)\,d\mu(s_i) = 1, \forall i = \overline{1,n}. \tag{7}$$

We make the following notations:

$$\Phi : U_0^n \to \mathbb{R},\ \Phi(g) = \int_{\mathbb{R}^n}\left[\mathcal{H}^\alpha(s)f^\alpha(s)\prod_{i=1}^n g_i^{1-\alpha}(s_i)\right]\,d\mu(s),$$
$$\Psi : \left(L^1(\mathbb{R})\right)^n \to \mathbb{R}^n,\ \Psi(g) = \left(\int_{\mathbb{R}} g_1(s_1)\,d\mu(s_1) - 1, \ldots, \int_{\mathbb{R}} g_n(s_n)\,d\mu(s_n) - 1\right)$$

Thus, (7) becomes

$$\min_{g\in U_0^n} \Phi(g),\ \Psi(g) = 0. \tag{8}$$

**LEMMA 1.** *$\Phi$ is a convex functional on $U_0^n$.*

*Proof.* It will suffice to show that the function $\varphi : \mathbb{R}^n \to \mathbb{R}$, $\varphi(\mathbf{x}) = (x_1 \cdot \ldots \cdot x_n)^{1-\alpha}$ is a convex one. For this, by observing that $\ln(\cdot)$ is concave, we prove that $\varphi(\cdot)$ is log-convex; for each $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and $t \in (0,1)$, one has:

$$\ln\left(\varphi\left[t\mathbf{x} + (1-t)\mathbf{y}\right]\right) = \ln\left(\prod_{i=1}^n \left[tx_i + (1-t)y_i\right]^{1-\alpha}\right) = (1-\alpha)\sum_{i=1}^n \ln\left[tx_i + (1-t)y_i\right] \leqslant$$
$$\leqslant (1-\alpha)\sum_{i=1}^n \left[t\ln x_i + (1-t)\ln y_i\right] = t\ln\left(\varphi(\mathbf{x})\right) + (1-t)\ln\left(\varphi(\mathbf{y})\right). \tag{9}$$

Now, $\varphi\left[t\mathbf{x} + (1-t)\mathbf{y}\right] \leqslant t\varphi(\mathbf{x}) + (1-t)\varphi(\mathbf{y})$ is equivalent with

$$\ln\left(\varphi\left[t\mathbf{x} + (1-t)\mathbf{y}\right]\right) \leqslant \ln\left[t\varphi(\mathbf{x}) + (1-t)\varphi(\mathbf{y})\right]. \tag{10}$$

Since $\varphi(\cdot)$ is log-convex, we have

$$\ln\left(\varphi\left[t\mathbf{x} + (1-t)\mathbf{y}\right]\right) \leqslant t\ln\left(\varphi(\mathbf{x})\right) + (1-t)\ln\left(\varphi(\mathbf{y})\right)$$

and, by concavity of $\ln(\cdot)$,

$$t\ln\left(\varphi(\mathbf{x})\right) + (1-t)\ln\left(\varphi(\mathbf{y})\right) \leqslant \ln\left(t\varphi(\mathbf{x}) + (1-t)\varphi(\mathbf{y})\right).$$

□

The Lagrange function of (8) is $L(g,\lambda)=\Phi(g)+\langle\lambda,\Psi(g)\rangle$.

**COROLLARY 1.** *For every $\lambda\in\mathbb{R}^n$, $L(\cdot,\lambda)$ is a convex function.*

*Proof.* Use Lemma 1 and the fact that $\Psi(\cdot)$ is affine. □

For every $h\in U^n$ and $g\in U_0^n$, it exists a $t_0>0$, such that $(g+t_0\cdot h)\in U_0^n$; $\Phi$ being convex, the function

$$t\to\frac{\Phi(g+t\cdot h)-\Phi(g)}{t}$$

is monotone, therefore the directional derivatives of the above operators can be calculated using a Lebesgue monotone convergence theorem:

$$D\Phi(g)(h)=\lim_{t\to 0}\frac{\Phi(g+t\cdot h)-\Phi(g)}{t}=(1-\alpha)\sum_{i=1}^{n}\int_{\mathbb{R}^n}\left[\mathcal{H}^\alpha(s)f^\alpha(s)g^{1-\alpha}(s)\frac{h_i(s_i)}{g_i(s_i)}\right]\,d\mu(s),$$

$$D\Psi(g)(h)=\lim_{t\to 0}\frac{\Psi(g+t\cdot h)-\Psi(g)}{t}=\left(\int_{\mathbb{R}}h_1(s_1)\,d\mu(s_1),\dots,\int_{\mathbb{R}}h_n(s_n)\,d\mu(s_n)\right),$$

$$DL(g,\overline{\lambda})(h)=D\Phi(g)(h)+\langle\overline{\lambda},D\Psi(g)(h)\rangle.$$

Using Theorem 3.4 from [1] $\overline{g}$ is solution to (8) if and only if it exists a $\overline{\lambda}\in\mathbb{R}^n$ such that

$$\overline{g}\in\operatorname{argmin}_{g\in U_\varepsilon^n}L(g,\overline{\lambda})\text{ and }\Psi(\overline{g})=0_{\mathbb{R}^n}.\tag{11}$$

For a given $\overline{\lambda}\in\mathbb{R}^n$, first condition in (11) is equivalent with $0\in\partial L(\overline{g},\overline{\lambda})$. As the $L(\cdot,\overline{\lambda})$ admits directional derivatives (at least) on every direction $h\in U^n$, a natural way to solve (11) is to find a solution to $DL(g,\overline{\lambda})(h)=0$, $\forall\,h\in U^n$ or, equivalently,

$$(\alpha-1)\sum_{i=1}^{n}\int_{\mathbb{R}^n}\left[\mathcal{H}^\alpha(s)f^\alpha(s)g^{1-\alpha}(s)\frac{h_i(s_i)}{g_i(s_i)}\right]\,d\mu(s)=\sum_{i=1}^{n}\overline{\lambda}_i\int_{\mathbb{R}}h_i(s_i)\,d\mu(s_i),\forall\,h\in U^n.\tag{12}$$

For an index $1\leqslant i\leqslant n$, and a vector $s=(s_1,s_2,\dots,s_n)\in\mathbb{R}^n$, let us denote

$$s^i=(s_1,\dots s_{i-1},s_{i+1},\dots,s_n).$$

**LEMMA 2.** *If $\overline{g}$ is a solution to (12), then, for every $1\leqslant i\leqslant n$, we have*

$$\overline{g}_i(s_i)=\frac{\displaystyle\int_{\mathbb{R}^{n-1}}\left[\mathcal{H}^\alpha(s)f^\alpha(s)\overline{g}^{1-\alpha}(s)\right]\,d\mu(s^i)}{\displaystyle\int_{\mathbb{R}^n}\left[\mathcal{H}^\alpha(s)f^\alpha(s)\overline{g}^{1-\alpha}(s)\right]\,d\mu(s)}\ a.\ e.,\ and\tag{13}$$

*if $X_i$ is a random variable having density $\overline{g}_i$, and $b:\mathbb{R}\to\mathbb{R}$ is a continuous function, then*

$$\mathbb{E}_{\overline{g}_i}\left[b(X_i)\right]=\frac{\displaystyle\int_{\mathbb{R}^n}\left[\mathcal{H}^\alpha(s)f^\alpha(s)\overline{g}^{1-\alpha}(s)b(s_i)\right]\,d\mu(s)}{\displaystyle\int_{\mathbb{R}^n}\left[\mathcal{H}^\alpha(s)f^\alpha(s)\overline{g}^{1-\alpha}(s)\right]\,d\mu(s)}.\tag{14}$$

*Proof.* equation (12) has the form:

$$\sum_{i=1}^{n}\int_{\mathbb{R}^n}A(s)h_i(s_i)\,d\mu(s)=\sum_{i=1}^{n}\int_{\mathbb{R}}\overline{\lambda}_ih_i(s_i)\,d\mu(s_i),\forall\,h\in U^n.$$

For a given $i$, let us choose $h_j\equiv 0$, for all $j\neq i$, and, denote

$$\Lambda_+^i=\left\{s_i\ :\int_{\mathbb{R}^{n-1}}A(s)\,d\mu(s^i)>\overline{\lambda}_i\right\},\ \Lambda_-^i=\mathbb{R}\setminus\Lambda_+^i.$$

We choose first $h_i(s_i) = \left[\int_{\mathbb{R}^{n-1}} A(s)\, d\mu(s^i) - \overline{\lambda}_i\right]^+$, second $h_i(s_i) = \left[\int_{\mathbb{R}^{n-1}} A(s)\, d\mu(s^i) - \overline{\lambda}_i\right]^-$, and we obtain

$$\int_{\Lambda_+^i} \left[\int_{\mathbb{R}^{n-1}} A(s)\, d\mu(s^i) - \overline{\lambda}_i\right]^2 d\mu(s_i) = 0 \text{ and } \int_{\Lambda_-^i} \left[\int_{\mathbb{R}^{n-1}} A(s)\, d\mu(s^i) - \overline{\lambda}_i\right]^2 d\mu(s_i) = 0$$

or, equivalently, $\int_{\mathbb{R}^{n-1}} A(s)\, d\mu(s^i) = \overline{\lambda}_i$, a. e. From (12) we get

$$g_i(s_i) = \frac{\alpha - 1}{\overline{\lambda}_i} \int_{\mathbb{R}^{n-1}} \left[\mathcal{H}^\alpha(s) f^\alpha(s) \overline{g}^{1-\alpha}(s)\right]\, d\mu(s^i).$$

Moreover, if, for any given $i$, we choose $h_i = \overline{g}_i$, and $h_j \equiv 0$ for all $j \neq i$, we get:

$$\overline{\lambda}_i = (\alpha - 1) \int_{\mathbb{R}^n} \left[\mathcal{H}^\alpha(s) f^\alpha(s) \overline{g}^{1-\alpha}(s)\right]\, d\mu(s).$$

Therefore, (13) follows easily; on the other hand (14) is an easy problem of calculation. □

# 3 Algorithm for estimating rare-event probabilities

In this section we suppose that $\mathcal{H}(\mathbf{X}) = \mathbb{1}_{[\mathcal{F}(\mathbf{X}) \geqslant a]}$, where $\mathcal{F}$ is a Lebesgue measurable function, $a \in R$, and $[\mathcal{F}(\mathbf{X}) \geqslant a]$ is a small probability event. We follow here the framework of a multistage procedure for the estimation of $\overline{g}$ - this type of algorithm appear often in the literature ([10], [2], [6]). The line of the algorithm is to build a sequence of thresholds $\left(\gamma^{(k)}\right)_{k \geqslant 0}$ which converge to $a$ and a sequence of densities $\left(g^{(k)}\right)_{k \geqslant 0}$ convergent to $\overline{g}$. The first sketch of the algorithm follows the equations (13):

---

deterministic version

**Step 1** $g^{(0)} \leftarrow f$; $k \leftarrow 1$;

**Step 2** do {

for $(i = \overline{1,n})$

$$(*)\ g_i^{(k)}(s_i) = \frac{\displaystyle\int_{\mathbb{R}^{n-1}} \left[\mathcal{H}^\alpha(s) f^\alpha(s) \left(g^{(k-1)}(s)\right)^{1-\alpha}\right]\, d\mu(s^i)}{\displaystyle\int_{\mathbb{R}^n} \left[\mathcal{H}^\alpha(s) f^\alpha(s) \left(g^{(k-1)}(s)\right)^{1-\alpha}\right]\, d\mu(s)};$$

$k++$;

}

until(some stopping conditions)

---

For continuous distributions clearly we can not implement this version of the algorithm, but, in most cases, using (14) we can completely determine the distribution $g^{(k)}$ using some of its moments (those who are sufficient for the calculation of expectation, variance etc). In such cases we change (*) by:

for $(r = \overline{1,p})$

$$\mu_i^{(k),r} = \frac{\displaystyle\int_{\mathbb{R}^n} \left[\mathcal{H}^\alpha(s) f^\alpha(s) \left(g^{(k-1)}(s)\right)^{1-\alpha} s_i^r\right]\, d\mu(s)}{\displaystyle\int_{\mathbb{R}^n} \left[\mathcal{H}^\alpha(s) f^\alpha(s) \left(g^{(k-1)}(s)\right)^{1-\alpha}\right]\, d\mu(s)};$$

and update $g^{(k)}$ using $\left(\mu_i^{(k),r}\right)_{\substack{i=\overline{1,n},\\ r=\overline{1,p}}}$;

In order to present the stochastic version of the above algorithm we make some notations - see also [7] and [6]. For a given level $\rho \in (0,1)$ we denote by $\gamma(g,\rho)$ a $(1-\rho)$-quantile of $\mathcal{F}(\mathbf{X})$ ($\mathbf{X}$ having density

$g$):

$$\mathcal{P}_g\left[\mathcal{F}(\mathbf{X}) \geqslant \gamma(g,\rho)\right] \geqslant \rho \text{ and } \mathcal{P}_g\left[\mathcal{F}(\mathbf{X}) \leqslant \gamma(g,\rho)\right] \geqslant 1-\rho$$

If $\mathbb{X} = (\mathbf{X}^1, \mathbf{X}^2, \ldots, \mathbf{X}^N)$ is a i. i. d. sample from $g$, then $\gamma(g,\rho)$ can be approximated by a $(1-\rho)$-sample quantile of $\mathcal{F}(\mathbf{X}^1), \mathcal{F}(\mathbf{X}^2), \ldots, \mathcal{F}(\mathbf{X}^N)$, denoted by $\gamma_N(\mathbb{X},\rho)$. In our algorithm we use the method from [6] and [7] which increase the sample size when the number of samples that is used for parameter updating becomes too small. We say that a given $\rho$ satisfy the condition $C(\rho)$ if $\gamma_N(\mathbb{X},\rho)$ is greater than or equal to $\min\{a, \gamma^{(k-1)} + \varepsilon\}$.

---

Stochastic version

**Step 1** initialization:

$g^{(0)} = f$; $k = 1$; $N$ an initial sample size; $\rho_0$ an initial quantile level;

generate i.i.d. samples $\mathbb{X} = (\mathbf{X}^1, \mathbf{X}^2, \ldots, \mathbf{X}^N)$ from $g^{(k-1)}$;

$\gamma^{(k)} = \min\{a, \gamma_N(\mathbb{X}, \rho_{k-1})\}$;

**Step 2** repeat until $(\gamma^{(k)} = a)$

- for $(i = \overline{1,n},\ r = \overline{1,p})$ do

$$(**)\ \mu_i^{(k),r} = \frac{\sum_{j=1}^{N}\left[\mathbb{1}_{\left[\mathcal{F}(\mathbf{X}^j)\geqslant\gamma^{(k)}\right]} f^{\alpha}(\mathbf{X}^j)\left(g^{(k-1)}(\mathbf{X}^j)\right)^{1-\alpha}\left(X_i^j\right)^r\right]}{\sum_{j=1}^{N}\left[\mathbb{1}_{\left[\mathcal{F}(\mathbf{X}^j)\geqslant\gamma^{(k)}\right]} f^{\alpha}(\mathbf{X}^j)\left(g^{(k-1)}(\mathbf{X}^j)\right)^{1-\alpha}\right]};$$

- update $g^{(k)}$ using $\left(\mu_i^{(k),r}\right)_{\substack{i=\overline{1,n},\\ r=\overline{1,p}}}$;
- if $C(\rho_{k-1})$ then $\rho_k = \rho_{k-1}$; $k++$; generate i.i.d. samples $\mathbb{X} = (\mathbf{X}^1, \mathbf{X}^2, \ldots, \mathbf{X}^N)$ from $g^{(k-1)}$ and update $\gamma^{(k)} \leftarrow \min\{a, \gamma_N(\mathbb{X}, \rho_{k-1})\}$;
- else, if $C(\rho')$ for some $\rho' < \rho_{k-1}$, then let $\rho_k$ be the largest of such $\rho'$ and $\gamma^{(k)} = \min\{a, \gamma_N(\mathbb{X}, \rho_{k-1})\}$;
- if $C(\rho')$ is not satisfied for any $\rho' \leqslant \rho_{k-1}$, then $N \leftarrow \nu \cdot N$; generate i.i.d. samples $\mathbb{X} = (\mathbf{X}^1, \mathbf{X}^2, \ldots, \mathbf{X}^N)$ from $g^{(k-1)}$ and update $\gamma^{(k)} = \min\{a, \gamma_N(\mathbb{X}, \rho_{k-1})\}$;

---

With final distribution $g^{(k)} \cong \overline{g}$ we estimate $m$ using (2):

$$m_N(g^{(k)}) = \frac{1}{N_0}\sum_{j=1}^{N_0} \mathbb{1}_{[\mathcal{F}(\mathbf{X}^j)\geqslant\gamma]}\frac{f(\mathbf{X}^j)}{g(\mathbf{X}^j)}.$$

An adaptation of the above algorithm can be used for optimization if in the initialization step we add $g^{(0)}, g^{(-1)} \leftarrow h_0$ ($h_)$ being an initial distribution), and change (**) by

$$(***)\quad \mu_i^{(k),r} = \frac{\sum_{j=1}^{N}\left[\mathbb{1}_{\left[\mathcal{F}(\mathbf{X}^j)\geqslant\gamma^{(k)}\right]}\left(g^{(k-2)}(\mathbf{X}^j)\right)^{\alpha}\left(g^{(k-1)}(\mathbf{X}^j)\right)^{1-\alpha}\left(X_i^j\right)^r\right]}{\sum_{j=1}^{N}\left[\mathbb{1}_{\left[\mathcal{F}(\mathbf{X}^j)\geqslant\gamma^{(k)}\right]}\left(g^{(k-2)}(\mathbf{X}^j)\right)^{\alpha}\left(g^{(k-1)}(\mathbf{X}^j)\right)^{1-\alpha}\right]};$$

# 4 Numerical results

In this section we present the numerical results of our algorithm on both estimation of rare events probabilities and continuous optimization. For small probabilities estimation we use a common example

from literature (see [2], [3], [10]). As for continuous optimization we illustrate the procedure on some well known benchmark problems ([2], [7]).

## 4.1 Probabilities of rare events

We test our algorithm using the example from [2] - a weighted bridge with random independent weights $X_i$, exponential distributed: $X_i \sim Exp(\lambda_i)$, $i = \overline{1,5}$. Consider the case $(\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5) = (1, 1, 3, 2, 10)$ and estimate the probability that the shortest path from $a$ to $b$ is at least $a \in \{7, 8, 9\}$.

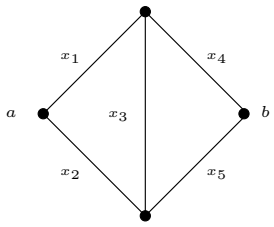



Figure 1: The bridge used as example

| $M$ | $a = 7$ | $a = 8$ | $a = 9$ |
|---|---|---|---|
| $1 \cdot 10^5$ | $6.20 \cdot 10^{-7}/0.021$ $(0.993s)$ | $6.62 \cdot 10^{-7}/0.022$ $(0.822s)$ | $7.90 \cdot 10^{-8}/0.007$ $(0.833s)$ |
| $2 \cdot 10^5$ | $5.34 \cdot 10^{-7}/0.069$ $(1.804s)$ | $5.76 \cdot 10^{-7}/0.022$ $(1.855s)$ | $9.20 \cdot 10^{-8}/0.011$ $(1.628s)$ |
| MC | $1.6 \cdot 10^{-7}/0.999$ $(815s)$ | $1.2 \cdot 10^{-7}/0.999$ $(754s)$ | $3 \cdot 10^{-8}/0.999$ $(761s)$ |

Table 1: Results for the estimation of various small probabilities

In all our tests the initial data are: sample size is $N = 1000$, quantile level $\rho_0 = 0.01$, sample size increasing factor is $\nu = 2$, $\varepsilon = 0.001$, and $\alpha = 1.5$. The results are listed in table 1 with $10^8$ samples for basic Monte Carlo; our algorithm estimates the probabilities using (for the final step) $M = 1 \cdot 10^5$ and $2 \cdot 10^5$ simulations replications. The algorithm is run for $n = 30$ different seeds, and table 1 presents the estimates, the relative errors and the execution times on average.

It can be seen that the time simulation effort is reduced by roughly a factor of 350. Our estimates are more accurate for smaller probabilities (i. e., for $a = 9$) and have smaller relative errors.

| | $k$ | $\gamma^{(k)}$ | $\mu_1^{(k),1}$ | $\mu_2^{(k),1}$ | $\mu_3^{(k),1}$ | $\mu_4^{(k),1}$ | $\mu_5^{(k),1}$ |
|---|---|---|---|---|---|---|---|
| $a = 7$ | 1 | 2.379 | 1.0 | 1.0 | 0.333 | 0.500 | 0.100 |
| | 2 | 5.895 | 2.518 | 2.742 | 0.086 | 0.199 | 0.012 |
| | 3 | 7.000 | 6.105 | 6.199 | 0.028 | 0.167 | 0.003 |
| $a = 8$ | 1 | 2.547 | 1.0 | 1.0 | 0.333 | 0.500 | 0.100 |
| | 2 | 6.311 | 2.920 | 2.880 | 0.260 | 0.376 | 0.107 |
| | 3 | 8.000 | 7.134 | 8.022 | 0.308 | 0.186 | 0.076 |
| $a = 9$ | 1 | 2.427 | 1.0 | 1.0 | 0.333 | 0.500 | 0.100 |
| | 2 | 6.311 | 2.497 | 2.550 | 0.160 | 0.269 | 0.075 |
| | 3 | 9.000 | 7.670 | 7.513 | 0.094 | 0.171 | 0.056 |

Table 2: Typical evolutions of the parameters

Table 2 shows the evolution of the means (first order moments) and the thresholds for different values of $a$. We see that the sequence of thresholds converges very fast; in almost all of the cases the desired value, $a$, is reached in the first three steps of the algorithm.

## 4.2 Continuous optimization

In the following experiments we use the family of multivariate normal distributions with independent components, that is, with probability density functions of product form. Initial mean vector has components randomly chosen from $[-50, 50]$ and covariance matrix is a diagonal one with variances 625. Other initial values are: sample size $N = 2000$, quantile level $\rho_0 = 0.01$ and $\alpha = 1.5$. Many of the problems from below are with constraints - in these cases we use the accept/reject method for generate appropriate samples. We consider global minimization problems having the form:

$$\min_{\mathbf{x}\in\chi} \mathcal{F}(\mathbf{x})$$

| | optimum | se | final $\gamma^{(k)}$ |
|---|---|---|---|
| $\mathcal{F}_1$ | 1.0001 | $6.9 \cdot 10^{-10}$ | 1.005 |
| $\mathcal{F}_2$ | 0.0043 | $6.3 \cdot 10^{-6}$ | 0.003 |
| $\mathcal{F}_3$ | 0.043 | $7.1 \cdot 10^{-5}$ | 0.0276 |
| $\mathcal{F}_4$ $(n = 2)$ | 0.0007 | $4.9 \cdot 10^{-7}$ | 0.0003 |
| $\mathcal{F}_4$ $(n = 10)$ | 7.1541 | $4.5 \cdot 10^{-3}$ | 7.0905 |
| $\mathcal{F}_5$ | 3.0000 | $4.7 \cdot 10^{-10}$ | 3.0000 |

Table 3: Results for different continuous multi-extremal optimization problems

The problems analyzed are enumerated below:

- a trigonometric function $(n = 10)$ which has frequent local minima:

$$\mathcal{F}_1(\mathbf{x}) = \sum_{i=1}^{n} 8\sin^2\left(7(x_i - 0.9)^2\right) + 6\sin^2\left(14(x_i - 0.9)^2\right) + (x_i - 0.9)^2,$$

  with $-10 \leqslant x_i \leqslant 10$, $\mathbf{x}^* = (0.9, 0.9, \ldots, 0.9)^t$, and $\mathcal{F}_1(\mathbf{x}^*) = 1$.

- Griewangk's function $(n = 10)$; it has many widespread local minima regularly distributed:

$$\mathcal{F}_2(\mathbf{x}) = 0.00025 * \sum_{i=1}^{n} x_i^2 - \prod_{i=1}^{n} \cos\left(\frac{x_i}{\sqrt{i}}\right) + 1,$$

  with $-600 \leqslant x_i \leqslant 600$, $\mathbf{x}^* = (0, 0, \ldots, 0)^t$, and $\mathcal{F}_2(\mathbf{x}^*) = 0$.

- Pinter's function $(n = 10)$ which has many local minima:

$$\begin{aligned}\mathcal{F}_3(\mathbf{x}) = &\sum_{i=1}^{n} i x_i^2 + 20\sum_{i=1}^{n} i\sin^2\left(x_{i-1}\sin x_i - x_i + \sin x_{i+1}\right) + \\ &+ \sum_{i=1}^{n} i\log_{10}\left[1 + i(x_{i-1}^2 - 2x_i + 3x_{i+1} - \cos x_i + 1)^2\right],\end{aligned}$$

  where $x_0 = x_n$, $x_{n+1} = x_1$, $\mathbf{x}^* = (0, 0, \ldots, 0)^t$, and $\mathcal{F}_3(\mathbf{x}^*) = 0$.

- Rosenbrock's valley $(n = 2, 10)$; global minimum lays inside a flat valley:

$$\mathcal{F}_4(\mathbf{x}) = \sum_{i=1}^{n-1} \left[100(x_{i+1} - x_i^2)^2 + (1 - x_i)^2\right],$$

  where $-2.048 \leqslant x_i \leqslant 2.048$, $\mathbf{x}^* = (1, 1, \ldots, 1)^t$, and $\mathcal{F}_4(\mathbf{x}^*) = 0$.

- Goldstein-Price's valley ($n = 2$):

$$\mathcal{F}_5(\mathbf{x}) = \left[1 + (x_1 + x_2 + 1)^2 \cdot \left(19 - 14x_1 + 3x_1^2 - 14x_2 + 6x_1x_2 + 3x_2^2\right)\right] \cdot \\ \cdot \left[30 + (2x_1 - 3x_2)^2 \cdot \left(18 - 32x_1 + 12x_1^2 + 48x_2 - 36x_1x_2 + 27x_2^2\right)\right],$$

where $-2.0 \leqslant x_i \leqslant 2.0$, $\mathbf{x}^* = (0, -1)$, and $\mathcal{F}_5(\mathbf{x}^*) = 3$.

Our algorithm estimates the optimum value using the final mean (i.e., first order moments: $\mu^{(k),1}$); the algorithm is run for $n = 30$ independent replications, and table 3 presents the functions values for the final means on average, corresponding standard errors, and (in parentheses) the final $\gamma^{(k)}$ values.

# References


[1] Barbu, V., T. Precupanu *Convexity and Optimization in Banach Spaces*, 4th edition, Springer Verlag, 2012.

[2] de Boer, P. T., D. P. Kroese, S. Mannor, R. Y. Rubinstein *A tutorial on the cross-entropy method*, Annals of Operations Research, 134, pp. 19-67, 2005.

[3] Botev, Z. I., D. P. Kroese *An efficient algorithm for rare-event probability estimation, combinatorial, and counting*, Methodology and Computing in Applied Probability, 10(4), pp. 471-505, 2008.

[4] van Erven, T., P. Harremoes, *Rényi divergence and majorization*, IEEE International Symposium on Information Theory Proceedings (ISIT), vol. 305, pp. 1335-1339, 2010.

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

[6] Homem-de-Mello, T., *A Study on the Cross-Entropy Method for Rare-Event Probability Estimation*, INFORMS Journal of Computing, vol. 19, no. 3, pp. 381-394, 2007.

[7] Hu, J., M. C. Fu, S. I. Marcus *A Model Reference Adaptive Search Method for Global Optimization*, INFORMS Operations Research, vol. 55, no. 3, pp. 549-568, 2007.

[8] Kroese, D. P., S. Porotsky, R, Y. Rubinstein, *The cross-entropy method for continuous multi-extremal optimization*, Methodology and Computing in Applied Probability, 8, pp. 383-407, 2006.

[9] Rubinstein, R. Y., *Optimization of computer simulation models with rare events*, Methodology and Computing in Applied Probability, 1, pp. 127-190, 1997.

[10] Rubinstein, R. Y., *The cross-entropy method for combinatorial and continuous optimization*, Methodology and Computing in Applied Probability, 1, pp. 127-190, 1999.