Minimum cross-entropy methods for rare-event simulation

Ad Ridder[1] Reuven Rubinstein

[1]*Department of Econometrics*
Vrije Universiteit Amsterdam
aridder@feweb.vu.nl

Rubinstein Conference at Sandbjerg Estate
July 14-18, 2008

CE and MCE

Cross-Entropy for rare-event simulation?!

There was already a CE-community at RESIM 2002, Madrid.

I got involved after a paper on CE for Markovian reliability systems.

Reuven visited me in Summer 2004.

He came with a new idea for rare-event simulation: Minimum Cross-Entropy (MinxEnt or MCE).

The core of CE for RESIM

Rare event problem

$$\ell = P(\boldsymbol{X} \in A).$$

Sometimes we write $P_f(\boldsymbol{X} \in A)$.

- $\boldsymbol{X}$ a random process of interest;
- f the statistical law or probability density of the process;
- A the rare event: ℓ is very small, say e-9 or less.

Importance sampling simulation using density g for estimating ℓ.

IS research is about 'how to find a good g?'.

Cross-Entropy

Program

Solve for density g

$$\inf \{D(g^*, g) : g \in \mathcal{G}\}.$$

- g^* is the optimal density (zero-variance), i.e., the original density f conditioned on the rare event:

$$g^*(\boldsymbol{x}) = f(\boldsymbol{x})\mathbf{1}\{\boldsymbol{x} \in A\}/\ell.$$

- $D(g^*, g)$ is the Kullback-Leibler cross-entropy or divergence

$$D(g^*, g) = \int g^*(\boldsymbol{x}) \log \frac{g^*(\boldsymbol{x})}{g(\boldsymbol{x})}\, d\boldsymbol{x}.$$

- $\mathcal{G}$ is a parameterized family of probability densities.

Reuven's new idea

Getting a good importance sampling density is about

1. control of the likelihood ratio;
2. make the rare event more likely;
3. generate samples with small variability.

Minimum Cross-Entropy

Program

Solve for density g

$$\inf D(g,f) \text{ s.t. one or more moment constraints.}$$

- f is the original density;
- $D(g,f)$ is the Kullback-Leibler cross-entropy or divergence

$$D(g,f) = \int g(\boldsymbol{x}) \log \frac{g(\boldsymbol{x})}{f(\boldsymbol{x})} \, d\boldsymbol{x}.$$

Note: nonparametric program, i.e., infinite-dimensional.

Historic notes

Maximizing (Shannon's) entropy or minimizing a cross-entropy under moment constraints go back to

Jaynes (1957 and 1963), and Kullback & Kairat (1966)

under the names of Maximum Entropy Principle, and Principle of Minimum Discrimination Information.

There is huge literature on applying entropy, cross-entropy, maximum entropy, minimum cross entropy (with and without constraints).

Wide area of research fields: information theory, Bayesian decision analysis, natural language processing, utility theory, computer vision, spatial physics, thermodynamics, statistical mechanics, statistical data analysis, etc.

Application to rare-event simulation seemed to be new.

How to solve the MCE program?

Suppose that the moment constraints are specified by equalities

$$\inf_{g \geq 0} \left\{ D(g,f) : \int g(\boldsymbol{x})\, d\boldsymbol{x} = 1,\ \int C_i(\boldsymbol{x}) g(\boldsymbol{x})\, d\boldsymbol{x} = c_i, i = 1, \ldots, m \right\}.$$

Leave out the non-negativity constraint, and apply the method of Lagrange multipliers. After some algebra,

The MCE solution

$$g(\boldsymbol{x}) = \frac{f(\boldsymbol{x})\ \exp\left(\sum_{i=1}^{m} \lambda_i C_i(\boldsymbol{x})\right)}{K(\boldsymbol{\lambda})}.$$

Here $K(\cdot)$ is the normalizing constant.

Notice:

- the solution is non-negative;
- the Lagrange multiplier λ_0 associated with the total mass constraint $\int g(\boldsymbol{x})\, d\boldsymbol{x} = 1$ cancels out;
- the other multipliers satisfy

$$\nabla \log K(\boldsymbol{\lambda}) = \boldsymbol{c};$$

- gives explicit expressions for the likelihood ratio:

$$\frac{f(\boldsymbol{x})}{g(\boldsymbol{x})} = K(\boldsymbol{\lambda})\, e^{-\sum_{i=1}^{m} \lambda_i C_i(\boldsymbol{x})}.$$

Illustrative example

Rare-event problem

$$\ell = P(S > \gamma),$$

where $S = X_1 + \cdots + X_n$ is the sum of a fixed number of i.i.d. rv's (increments), and rarity parameter $\gamma \to \infty$.

MCE program

$\inf_g D(g, f)$ s.t.

(i) $E_g[S] = \gamma$ (make the rare event more likely);

(ii) $E_g[S^2] = \gamma^2 + \epsilon$ (small variability in the generated samples).

Hence $\mathrm{Var}_g[S] = \epsilon$ meaning that most samples of S stay close to the overflow level γ.

It works nicely!?

Let $X_1, X_2, \ldots, X_n \overset{f}{\sim} N(\mu, \sigma^2)$ i.i.d.

- Solution to the MCE program with only the first moment constraint is

$$X_1, X_2, \ldots, X_n \overset{h}{\sim} N(\gamma/n, \sigma^2) \quad \text{i.i.d.}$$

- Solution to the MCE program with both moment constraints is a multivariate (correlated) normal distribution:

$$\boldsymbol{X} = (X_1, X_2, \ldots, X_n) \overset{g}{\sim} N((\gamma/n)\boldsymbol{e}, \Sigma).$$

The partial sum is normal in all these three cases:

$$S \overset{f}{\sim} N(n\mu, n\sigma^2); \quad S \overset{h}{\sim} N(\gamma, n\sigma^2); \quad S \overset{g}{\sim} N(\gamma, \epsilon).$$

Result

Denote $Y = L1\{S > \gamma\}$ for the IS estimator
where L represents the likelihood ratio.

Performance in terms of relative error of the estimator:

$$\mathrm{RE}[Y] = \frac{E[Y^2]}{(E[Y])^2}.$$

Set $\epsilon = n\sigma^2/\kappa$ for $\kappa > 1$, then

$$\mathrm{RE}_g[Y] \approx \frac{1}{\sqrt{\kappa}}\,\mathrm{RE}_h[Y].$$

We could prove for $\kappa \leq 2$ only
empirically we found these improvements for larger κ upto some κ_0 (dependent on the parameters).

A trivial counter example

Consider $\ell = P(X > \gamma)$ when $X \overset{f}{\sim} \mathrm{Exp}(1)$.

Solution to the MCE program (with both moment constraints)

$$g(x) = \frac{e^{-x}\, e^{\lambda_1 x + \lambda_2 x^2}}{K(\lambda_1, \lambda_2)} \quad (x > 0),$$

where λ_1, λ_2 are the Lagrange multipliers, and the normalizing constant $K(\lambda_1, \lambda_2) < \infty$ iff $\lambda_2 < 0$.
Consider the second moment of IS estimator Y:

$$\begin{aligned} \int_\gamma^\infty \left(\frac{f(x)}{g(x)}\right)^2 g(x)\, dx &= \int_\gamma^\infty \frac{f(x)}{g(x)} f(x)\, dx \\ &= K(\lambda_1, \lambda_2) \int_\gamma^\infty e^{-(\lambda_1+1)x - \lambda_2 x^2}\, dx = \infty. \end{aligned}$$

What went wrong

Controlling the likelihood ratio is the key issue in MCE.

What equality or inequality constraints might be appropriate?

Some MCE programs with a single constraint

1. As earlier: $\ell = P(S_n = X_1 + \cdots + X_n > \gamma)$, with i.i.d. light-tailed increments, n fixed, $\gamma \to \infty$.

$$\inf_{g \geq 0} \left\{ D(g, f) : \int g(\boldsymbol{x})\, d\boldsymbol{x} = 1,\ E_g[S_n] = \gamma \right\}.$$

Solution g factorizes, $X_1, \ldots, X_n$ remain i.i.d., with an exponentially tilted version of the original (marginal) density:

$$g(x) = f(x) \exp(\lambda x) / \text{normalizing constant.} \qquad (1)$$

Notice $E_g[X_j] = \gamma/n$ for all increments.

2. In the above, suppose that $\gamma = \gamma_n = na$, with $E_f[X_j] < a$.

   The ℓ_n satisfy a large deviations as $n \to \infty$.

   The MCE solution (1) coincides with the classic IS density obtained by the 'optimal path' heuristic from this LD.

   Its associated estimator is asymptotically optimal.

3. Now suppose that

$$\ell_n = P(S_n \leq -na(1+\epsilon) \text{ or } S_n \geq na),$$

where $a > 0$ and $\epsilon > 0$. And suppose that for the large deviations rate function $J(a) < J(-a(1+\epsilon))$.

This is the famous counter example to the LD approach of 2., see Glassermann & Wang (1997) or Bucklew (2004).

In case of standard Gaussian increments consider the MCE program

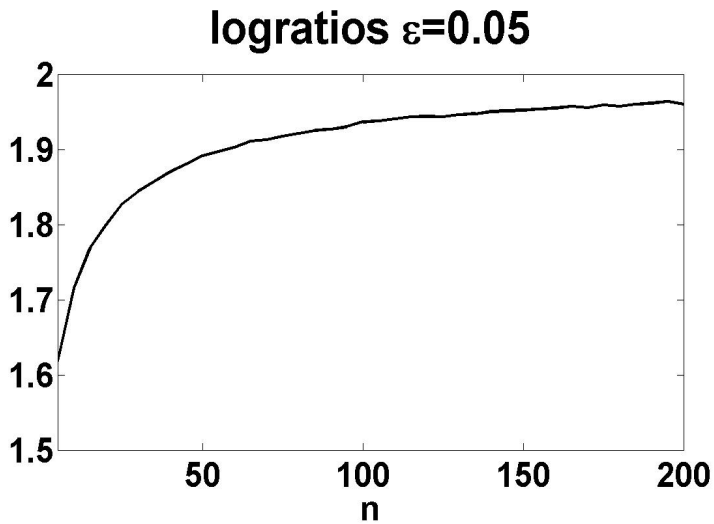
$$\inf_{g \geq 0} \left\{ D(g,f) : \int g(\boldsymbol{x})\, d\boldsymbol{x} = 1,\ E_g[S_n^2] = \rho n^2 \right\}.$$

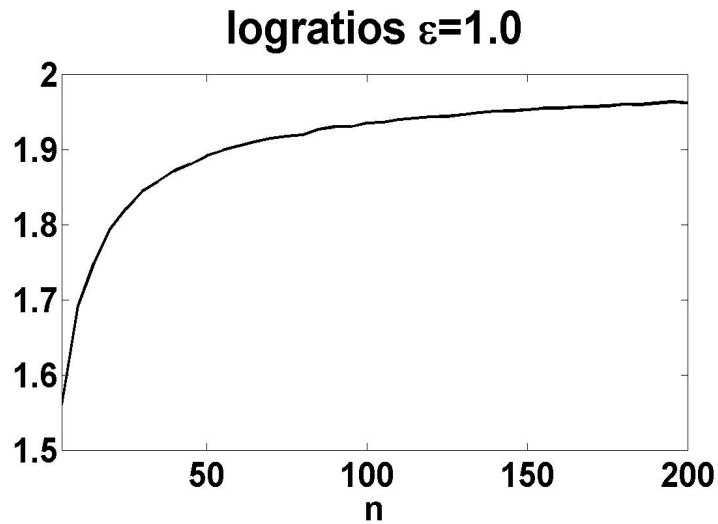
Parameter $\rho > 1/n$ is constant, e.g. $\rho = 1$, or proportional to n, e.g. $\rho = n/4$. We can prove for the associated IS estimator Y_n:

$$\liminf_{n \to \infty} \frac{\log E_g[Y_n^2]}{\log E_g[Y_n]} \geq \frac{2}{(1+\epsilon)^2}.$$

(NB: more about this ‘two-sided events problem’ in the next talk by Thomas Taimre).

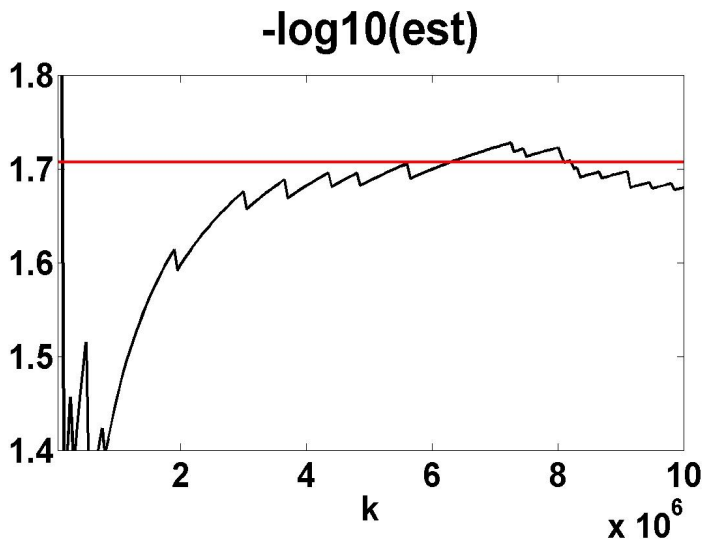
Experiments with fixed sample size and varying n.

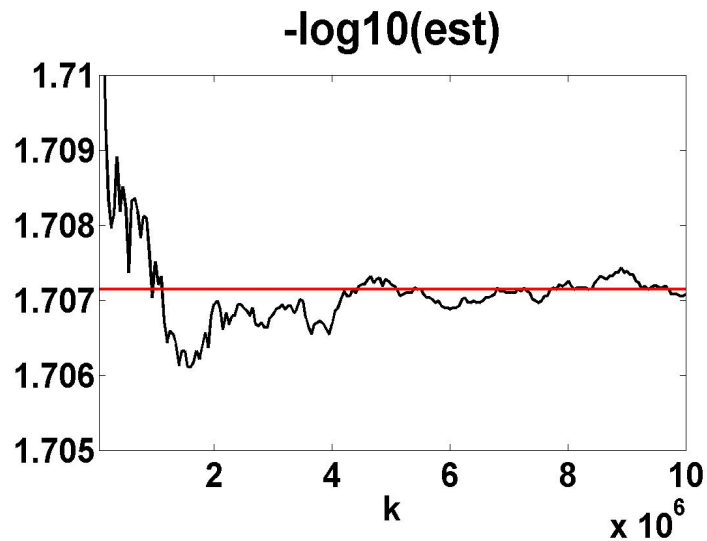




Experiments with varying sample size and constant n.

Left: Large deviations solution. Right: MCE solution.





4. Again $\ell = P(S = X_1 + \cdots + X_n > \gamma)$, but with i.i.d. heavy-tailed increments, n fixed, $\gamma \to \infty$.

   To be specific: subexponential X_j's with a concave hazard rate function $Q(x) = -\log P(X_j > x)$.

   Examples: Weibull (with shape parameter < 1), Pareto, Lognormal.

MCE program:

$$\inf_{g \geq 0} \left\{ D(g,f) : \int g(\boldsymbol{x})\, d\boldsymbol{x} = 1,\ E_g\Big[\sum\nolimits_{j=1}^{n} Q(X_j) \Big] = c \right\}.$$

Solution g factorizes, $X_1, \ldots, X_n$ remain i.i.d., with an hazard rate twisted version of the original (marginal) density (Juneja & Shahabuddin 2002):

$$g(x) = f(x) \exp(\lambda Q(x)) / \text{normalizing constant.} \qquad (2)$$

Optimal version obtained with the right choice $c = Q(\gamma)$.

5. However, consider

$$\inf_{g \geq 0} \left\{ D(g,f) : \int g(\boldsymbol{x})\, d\boldsymbol{x} = 1,\ E_g\Big[Q\Big(\sum_{j=1}^{n} X_j \Big) \Big] = c \right\}.$$

Solution gives correlated increments:

$$g(\boldsymbol{x}) = f(\boldsymbol{x}) \exp(\lambda Q(S)) / \text{normalizing constant.} \qquad (3)$$

This correlated hazard rate twisted density has heavier tails than its independent counter part (2) of previous slide:

$$E\Big[Q\Big(\sum_{j=1}^{n} X_j \Big) \Big] \leq E\Big[\sum_{j=1}^{n} Q(X_j) \Big].$$

Statistical result

Set constraint RHS $c = \rho Q(\gamma)$ with $0 < \rho < 1$ arbitrary.

Theorem

The importance sampling estimator Y using g of (3) is asymptotically optimal (as $\gamma \to \infty$).

The proof is based on

(i) let $\lambda = \lambda(\rho, \gamma)$ be the Lagrange multiplier in the solution (3);

(ii) show that $\lambda \uparrow 1$ as $\gamma \to \infty$, for any $0 < \rho < 1$;

(iii) show that the logratio $\log E_g[Y^2] / \log E_g[Y]$ is asymptotically at least $1 + \lambda$ as $\gamma \to \infty$.

Generating samples

Recall: g is multivariate of dimension n and does not factorize.
We have analysed three algorithms:

1. Acceptance-rejection.
2. Metropolis-Hastings.
3. Gibbs sampler.

Empirical diagnostics for testing on

(i) convergence; (ii) dependency structure; (iii) stationarity.

Overall the Gibbs sampler performed best.

Empirical results

We have executed numerous experiments and compared the performance of our importance sampling estimator (RR) with those obtained by implementing the JS algorithms (Juneja & Shahabuddin 2002: independent hazard rate twisted versions), and the AK algorithms (Asmussen & Kroese 2006: conditional Monte Carlo).

As expected, RR improves JS, but is outperformed (in most cases) by AK.
RR performs best in case of heavy, but not too heavy tails, e.g. Weibull with shape parameter $0.75 < \beta < 1$.

Discussion

MCE has its limitations for applying to rare-event simulation.

There are a few ‘success stories’ to report.

Further investigations are on the way, for instance the level crossing problem $P(X_1 + \cdots + X_N > \gamma)$ with a random sum of i.i.d. increments, and dynamic or sequential MCE (next talk by Thomas Taimre), parametric MCE.