The Minimum Cross Entropy Method For Rare Event Simulations

Ad Ridder
Econometrics
Vrije Universiteit, Amsterdam, Netherlands

Reuven Rubinstein
Faculty of Industrial Engineering and Management
Technion, Haifa, Israel

May 24, 2005

Abstract

This paper describes a new idea of finding the importance sampling density in rare events simulations: the MinxEnt method (shorthand for minimum cross-entropy). Some preliminary results show that the method might be very promising.

1 The minxent program

Assume

- $\boldsymbol{X} = (X_1, \ldots, X_n)$ is a random vector (with values denoted by $\boldsymbol{x}$);
- h is the joint density function of $\boldsymbol{X}$;
- $S_j(\cdot)$ $(j = 1, \ldots, k)$ are functions of $\boldsymbol{x}$;

Recall the Kullback-Leibler distance between any two density functions f, h of $\boldsymbol{X}$:

$$\mathcal{D}(f|h) = \mathbb{E}_f \left[\log \frac{f(\boldsymbol{X})}{h(\boldsymbol{X})} \right],$$

when $f \ll h$ (otherwise $\mathcal{D}(f|h) = \infty$). Consider the minxent optimisation program for finding a density f that minimises the KL-distance subject to constraints [4]:

$$\begin{aligned} &\min_f \mathcal{D}(f|h) \\ &\text{s.t. } \mathbb{E}_f[S_j(\boldsymbol{X})] = \theta_j, \quad j = 1, \ldots, k, \\ &\int f(\boldsymbol{x})\, d\boldsymbol{x} = 1. \end{aligned}$$

The solution of the MinxEnt program is [4]

$$f(\boldsymbol{x}) = \frac{h(\boldsymbol{x})\, e^{-\sum_{j=1}^k \lambda_j S_j(\boldsymbol{x})}}{c(\boldsymbol{\lambda})}, \tag{1}$$

where the λ_j's are Lagrange multipliers satisfying (in vector notation)

$$\nabla \log c(\boldsymbol{\lambda}) = -\boldsymbol{\theta}, \tag{2}$$

and $c(\cdot)$ is the normalising constant:

$$c(\boldsymbol{\lambda}) = c(\lambda_1, \ldots, \lambda_k) = \mathbb{E}_h \exp(-\sum_{j=1}^k \lambda_j S_j(\boldsymbol{X})). \tag{3}$$

2 Application to Rare Event Simulation

Let $X_1, X_2, \ldots, X_n$ i.i.d. with common density function $h_1(x)$, i.e., the joint density function is the product:

$$h(\boldsymbol{x}) = \prod_{i=1}^{n} h_1(x_i).$$

We denote the generic marginal by X with density h_1 which we also denote by h when there is no confusion. In this paper we assume that the marginal density h is light-tailed.

We shall consider the following type of rare events, also called a large buffer type of rare event:

$$A_n(\gamma) = \left\{ \sum_{i=1}^{n} X_i > \gamma \right\},$$

where n is constant. The rarity parameter is the level γ which we assume $\gamma \to \infty$. The problem is to estimate the rare event probability $\mathbb{P}_h(A_n(\gamma))$. Notice that we add the extra information of the density of the involved random variables, because we will apply other densities. The rare event probability is estimated by simulation using importance sampling, where the new density of the vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ is given by f which follows from solving a minxent program.

The importance sampling estimator becomes

$$Z_n(\gamma) = L_n(\boldsymbol{X}, h, f) 1\{A_n(\gamma)\},$$

where $L_n(\boldsymbol{X}, h, f)$ is the likelihood ratio of generating the vector $\boldsymbol{X}$ using the new density f given the original model density (or prior density) h:

$$L_n(\boldsymbol{X}, h, f) = \frac{h(\boldsymbol{X})}{f(\boldsymbol{X})} = \frac{\prod_{i=1}^{n} h_1(X_i)}{f(\boldsymbol{X})},$$

where we allow that under the new measure the X_i's are not i.i.d. anymore.

Our main focus is the efficiency of the IS estimator. Consider the squared coefficient of variation (SCV), or the squared relative error of the estimator:

$$\kappa_n^2(\gamma) = \frac{\mathbb{V}\mathrm{ar}_f[Z_n(\gamma)]}{(\mathbb{E}_f[Z_n(\gamma)])^2}.$$

Then $Z_n(\gamma)$ is logarithmically efficient, or asymptotically optimal, if $\kappa_n^2(\gamma)$ is at most a polynomial expression of the rarity parameter [1]. Equivalently, when

$$\frac{\mathbb{V}\mathrm{ar}_f[Z_n(\gamma)]}{(\mathbb{E}_f[Z_n(\gamma)])^{2-\epsilon}}$$

is bounded for any $\epsilon > 0$.

3 The Sanov solution

Let

$$S(\boldsymbol{X}) = \sum_{i=1}^{n} X_i,$$

and consider the minxent program with the following single constraint

$$\min_f \left\{ \mathcal{D}(f|h) \,:\, \mathbb{E}_f\left[S(\boldsymbol{X})\right] = \gamma \right\}. \tag{4}$$

It forces the new density to generate samples that lie most likely in or close to the rare event $A_n(\gamma)$. Under the optimal joint density $f(\boldsymbol{x})$ the variables $X_1, \ldots, X_n$ are i.i.d. with common density

$$f_1(x) = \frac{h_1(x)\exp(-\lambda x)}{\mathbb{E}_h[\exp(-\lambda X)]}, \tag{5}$$

where the Lagrange multiplier λ solves

$$\frac{\mathbb{E}_h[X\exp(-\lambda X]}{\mathbb{E}_h[\exp(-\lambda X)]} = \gamma/n \tag{6}$$

(this follows after some standard manipulations involving the normalising constant (3) and using independence). We obtain specifically for the generic marginal

$$\mathbb{E}_f[X] = \gamma/n.$$

And we notice that the optimal marginal density is an exponentially tilted density with tilting parameter $-\lambda$.

This approach of solving the MinxEnt program with the single constraint relates to applying large deviation theory to derive the optimal tilting parameters for rare events with light-tailed distributions [2]. For that matter we have to consider both $\gamma \to \infty$ and $n \to \infty$ proportionally

$$\gamma = \alpha n.$$

When we apply Sanov's theorem [3] we get

$$\lim_{n\to\infty} \frac{1}{n} \log P_h\left(\frac{1}{n}\sum_{i=1}^{n} X_i > \alpha\right) = -\inf_{f_1}\{\mathcal{D}(f_1|h_1) : \mathbb{E}_{f_1}[X_1] = \alpha\}, \tag{7}$$

and

$$\lim_{\gamma\to\infty} \frac{1}{\gamma} \log P_h\left(\sum_{i=1}^{n} X_i > \gamma\right) = \frac{1}{\alpha}\lim_{n\to\infty}\frac{1}{n}\log P_h\left(\frac{1}{n}\sum_{i=1}^{n} X_i > \alpha\right).$$

The rightside in (7) is exactly a minxent program, viz. for the i.i.d. marginals. Therefore we call the tilting density (5) the Sanov solution of the importance sampling.

3.1 Examples

- **Example A.**
  The X_i's are Bernoulli distributed according to

  $$\mathbb{P}_h(X = 1) = p, \quad \mathbb{P}_h(X = -1) = 1 - p,$$

  where $0 < p < 1$. Notice that $\mathbb{E}_h[X] = 2p - 1$. Solving (6) for the Lagrange multiplier yields (after some algebra)

  $$e^{2\lambda} = \frac{p}{1-p}\,\frac{1-\gamma/n}{1+\gamma/n}.$$

  The marginal X is again Bernoulli distributed:

  $$\mathbb{P}_f(X = 1) = (1+\gamma/n)/2, \quad \mathbb{P}_f(X = -1) = (1-\gamma/n)/2.$$

- **Example B.**
  The X_i's are normally distributed $\sim \mathcal{N}(\mu, \sigma^2)$. Solving (6) for the Lagrange multiplier yields (after some algebra)
  $$\mu - \lambda\sigma^2 = \gamma/n \;\Leftrightarrow\; \lambda = \frac{\mu - \gamma/n}{\sigma^2}.$$
  The marginal X is again normally distributed:
  $$X \sim f = \mathcal{N}(\mu - \lambda\sigma^2, \sigma^2) = \mathcal{N}(\gamma/n, \sigma^2).$$

- **Example C.**
  The X_i's are exponentially distributed with parameter u^{-1}, that is $\mathbb{E}_h[X] = u$, or
  $$h_1(x) = \frac{1}{u} e^{-x/u}.$$
  Type 1 rare event.
  Solving (6) for the Lagrange multiplier yields
  $$\lambda = \frac{n}{\gamma} - \frac{1}{u}.$$
  And the marginal X has new exponential density f_1 with parameter $n\gamma^{-1}$, i.e., $\mathbb{E}_f[X] = \gamma/n$.

4 Adding a variance constraint

In Sections 2 and 3 we considered the problem of estimating $\mathbb{P}_h(S(\boldsymbol{X}) > \gamma)$, where $S(\boldsymbol{X})$ is the partial sum of i.i.d. random variables, and we proposed an importance sampling simulation with new density f being the Sanov solution for which the first moment satisfies $\mathbb{E}_f[S(\boldsymbol{X})] = \gamma$. The idea is to generate samples that lie on average in or close to the rare event. The next step would be reducing the variance of $S(\boldsymbol{X})$ to force more samples staying close to their mean, and hence more samples reach the rare event.

4.1 The variance of the Sanov solutions

First, let us investigate what the variances are for the Sanov solutions in the three examples introduced in section 3.1.

- **Example A.**
  In this example the rare event appears when the rarity parameter $\gamma \uparrow n$ where n is fixed. Therefore, say $\gamma = n - (1/k)$ and $k \to \infty$. Because $\mathbb{E}_f[X] = \gamma/n$, we get $\mathbb{V}\mathrm{ar}_f[X] = 1 - (\gamma/n)^2$, and
  $$\begin{aligned}\mathbb{V}\mathrm{ar}_f[S(\boldsymbol{X})] &= n\left(1 - \left(\frac{\gamma}{n}\right)^2\right) = n\left(1 - \left(1 - \frac{1}{nk}\right)^2\right) \\ &= \frac{1}{k}\left(2 - \frac{1}{nk}\right),\end{aligned}$$
  where $k \to \infty$. Thus already the Sanov solution has a small variance.

- **Example B.**
  Because the terms of the partial sum $S(\boldsymbol{X})$ are i.i.d. normally distributed with mean γ/n and variance σ^2 (for the new density f), the partial sum itself has variance $n\sigma^2$, which is a constant (recall n is fixed, $\gamma \to \infty$). Hence it might be worthwhile to add a variance restriction to the minxent program.

- **Example C.**
  In this case the individual terms are i.i.d. exponentially distributed with parameter $n\gamma^{-1}$ (under the new density f). Thus
  $$\mathbb{V}\mathrm{ar}_f[S(\boldsymbol{X})] = n\left(\frac{\gamma}{n}\right)^2 = \tfrac{1}{n}\gamma^2.$$
  And here it would be really helpful to impose a restriction to the variance in the minxent program.

Conclusion: it depends on the distributions involved whether it would be worthy considering a variance constraint in the minxent program.

4.2 The double constraint minxent

In this section we add a variance constraint to the minxent program by considering the second moment. This, because the first constraint fixes the first moment, see (4). The new program becomes

$$\min_f \Big\{\mathcal{D}(f|h) \,:\, \mathbb{E}_f\left[S(\boldsymbol{X})\right] = \gamma,\ \mathbb{E}_f\left[(S(\boldsymbol{X}))^2\right] = \zeta\Big\}, \tag{8}$$

where

$$\zeta \geq \gamma^2. \tag{9}$$

A solution to this minxent program (called a minxent-2 solution) forces the new density to generate samples that lie most likely in or close to the rare event $A_n(\gamma)$ and it controls the variability of the samples around their mean. The solution is (cf. section 1):

$$f(x_1,\ldots,x_n) = \frac{h(x_1,\ldots,x_n)\exp(-\lambda_1(x_1+\cdots+x_n) - \lambda_2(x_1+\cdots+x_n)^2)}{c(\lambda_1,\lambda_2)}.$$

The normalising constant is

$$c(\lambda_1,\lambda_2) = \mathbb{E}_h[\exp(-\lambda_1 S - \lambda_2 S^2)].$$

The multipliers λ_1 and λ_2 follow from (2):

$$\begin{cases} \frac{\partial}{\partial\lambda_1}\log c(\lambda_1,\lambda_2) = -\gamma \\ \frac{\partial}{\partial\lambda_2}\log c(\lambda_1,\lambda_2) = -\zeta. \end{cases} \tag{10}$$

4.3 The double constraint solutions in the examples

We concentrate on the previous examples. Example A (the Bernoulli case) is dropped.

- **Example B.** (The normal case)

  Lemma 1. *The minxent-2 solution gives $X_1, X_2, \ldots, X_n$ to have a multivariate normal distribution with mean γ/n for each marginal and covariance matrix Σ for which*

  $$\Sigma_{ii} = \sigma^2\left(1 - \frac{2\lambda_2\sigma^2}{1+2n\lambda_2\sigma^2}\right), \quad \Sigma_{ij} = \sigma^2\frac{-2\lambda_2\sigma^2}{1+2n\lambda_2\sigma^2},$$

  for all i and $j \neq i$.

  The lagrange multipliers λ_1 and λ_2 solve the normalising equations (10). Hence, $S(\boldsymbol{X})$ has a normal distribution with mean γ and variance

  $$n\Sigma_{11} + n(n-1)\Sigma_{12} = \zeta - \gamma^2.$$

- **Example C.** (The exponential case)

  Lemma 2. *The minxent-2 solution gives $X_1, X_2, \ldots, X_n$ that satisfy: for each marginal variable X_k, given values $x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_n$ of the other marginals, its density (conditioned on these other values) is a conditional normal distribution conditioned on positive outcomes.*

The proofs of these two lemma's are obtained by careful calculus.

5 Asymptotically optimal importance sampling

In this section we investigate the asymptotical optimality of the importance sampling estimator for rare events of type 2 (type 1 is familiar and large deviations deal with it). Recall that this estimator is $Z = L1\{S(\boldsymbol{X}) > \gamma\}$, where $S(\boldsymbol{X}) = \sum_{i=1}^{n} X_i$ for i.i.d. X_i's, n is fixed, and the rarity parameter $\gamma \to \infty$. We have to find asymptotics for its SCV

$$\kappa^2(\gamma) = \frac{\mathbb{V}\mathrm{ar}_f[Z]}{(\mathbb{E}_f[Z])^2} = \frac{\mathbb{E}_f[Z^2]}{(\mathbb{E}_f[Z])^2} - 1.$$

Again we focus on the normal and exponential cases, and again the proofs are not given since these are mainly tedious algebra.

5.1 The normal case

Our first result concerns the Sanov solution of section 3.

Lemma 3. *When f is the Sanov solution the SCV is $O(\gamma)$ as $\gamma \to \infty$, more precisely*

$$\kappa^2(\gamma) \approx \frac{\gamma - n\mu}{2\sigma\sqrt{n}} - 1, \tag{11}$$

Now let us see whether the minxent-2 solution can improve this. We impose the variance of the minxent-2 solution to be less than the Sanov solution which has variance $n\sigma^2$, thus the righthand side ζ in the second constraint should be, cf. (9),

$$\gamma^2 \le \zeta < \gamma^2 + n\sigma^2.$$

The lower bound gives $S(\boldsymbol{X})$ zero variance.

Lemma 4. *Parameterise ζ by*

$$\zeta = \gamma^2 + \frac{n\sigma^2}{\rho}, \quad \rho > 1.$$

Then, when using the minxent-2 solution, the SCV of the importance sampling estimator is $O(\gamma)$ as $\gamma \to \infty$. Furthermore, for $\rho \le 2$

$$\kappa^2(\gamma) \approx \frac{\gamma - n\mu}{2\sigma\sqrt{n}\sqrt{\rho}} - 1. \tag{12}$$

Comparing the SCV (11) when the Sanov solution is used and (12) when the minxent-2 solution is used, we observe an improvement by a factor $1/\sqrt{\rho}$ where $\rho > 1$. We have shown this analytically for ρ up to value 2, however experiments show that the SCV is improved for larger values up to some ρ_0 dependent on the parameters (see section 6).

5.2 The exponential case

First a result concerning the Sanov solution of section 3.

Lemma 5. *When f is the Sanov solution the SCV is $O(\gamma)$ as $\gamma \to \infty$,*

Now let us see whether the minxent-2 solution can improve this. We impose the variance of the minxent-2 solution to be less than the Sanov solution which has variance γ^2/n, thus the righthand side ζ in the second constraint should be, cf. (9),

$$\gamma^2 \le \zeta < \gamma^2 + \gamma^2/n.$$

The lower bound gives $S(\boldsymbol{X})$ zero variance.

Lemma 6. *Parameterise ζ by*

$$\zeta = \gamma^2 + \frac{\gamma^2/n}{\rho}, \quad \rho > 1.$$

Then, when using the minxent-2 solution, the SCV of the importance sampling estimator improves the SCV of the Sanov solution.

6 Experiments

We compare the SCV of the importance sampling estimators when using Sanov and minxent-2 solutions which we denote by the densities g (Sanov) and f (minxent-2).

- **Example B: the normal case**
  Recall:
  $$S(\boldsymbol{X}) \overset{h}{\sim} \mathcal{N}(n\mu, n\sigma^2), \;\; S(\boldsymbol{X}) \overset{g}{\sim} \mathcal{N}(\gamma, n\sigma^2), \;\; S(\boldsymbol{X}) \overset{f}{\sim} \mathcal{N}(\gamma, n\sigma^2/\rho).$$
  Let $\mu = 5, \sigma^2 = 3$, i.e., $\mathbb{E}_h[X] = 5$.

  All importance sampling simulations are executed with sample size 1000. The experiments show that the SCV is convex unimodal as a function of $\rho > 1$, that is, the SCV decreases until some ρ_0 and increases thereafter.

  1. $n = 1, \gamma = 20$.
     Probability estimate: 2.632182e-018
     SCV κ_g^2 of Sanov solution: 0.009438
     Table of SCV κ_f^2 of minxent-2 solution for increasing ρ:

ρ	1.5	2	5	10	20	50
κ_f^2	0.007539	0.006406	0.003744	0.002774	0.001713	0.001912

  2. $n = 5, \gamma = 50$.
     Probability estimate: 6.086148e-011
     SCV κ_g^2 of Sanov solution: 0.007020
     Table of SCV κ_f^2 of minxent-2 solution for increasing ρ:

ρ	1.5	5	10	20	50	100
κ_f^2	0.005528	0.002988	0.001860	0.001294	0.001352	0.002118

  3. $n = 10, \gamma = 100$.
     Probability estimate: 3.512451e-020
     SCV κ_g^2 of Sanov solution: 0.010967
     Table of SCV κ_f^2 of minxent-2 solution for increasing ρ:

ρ	1.5	5	25	50	100	150
κ_f^2	0.008755	0.004443	0.001605	0.001329	0.002747	0.004888

  It is attractive to find the optimal scale factor ρ by the cross entropy method [5]:
  $$\rho^* = \arg\max_{\rho} \mathbb{E}_h[1\{S > \gamma\} \log f_\rho(S)],$$
  where $f_\rho = \mathcal{N}(\gamma, n\sigma^2/\rho)$. This optimal ρ is approximated iteratively: starting with some ρ_0 we get $\rho_1, \rho_2, \ldots$ according to
  $$\rho_{k+1} = \arg\max_{\rho} \mathbb{E}_{\rho_k}[L(S, h, f_{\rho_k}) 1\{S > \gamma\} \log f_\rho(S)],$$
  where the likelihood ratio
  $$L(x, h, f_{\rho_k}) = \frac{h(x)}{f_{\rho_k}(x)}.$$
  Working out the first order condition we get
  $$\rho_{k+1} = n\sigma^2 \frac{\mathbb{E}_{\rho_k}[L(S, h, f_{\rho_k}) 1\{S > \gamma\}}{\mathbb{E}_{\rho_k}[L(S, h, f_{\rho_k}) 1\{S > \gamma\}(S - \gamma)^2}.$$
  For instance for $n = 5$ in the experiments above we found starting with $\rho_0 = 2$ after 2-5 iterations that $\rho^* \approx 27$. Indeed, for this scale factor the SCV was 0.001109 (compare with the other values in the table above). And $n = 10$ gave $\rho^* \approx 50$.

- **Example C: the exponential case**
  Recall that under the original h and Sanov g the $X_1, X_2, \ldots$ are i.i.d. with marginals

$$X \overset{h}{\sim} \text{Exp}(u^{-1}), \quad X \overset{g}{\sim} \text{Exp}(\gamma^{-1}).$$

  For minxent-2, $X_1, \ldots, X_n$ are correlated with conditional marginal densities:

$$X_k|(X_j, j \neq k) \overset{f}{\sim} \mathcal{N}(\mu, \sigma^2) \quad \text{conditional on } \; x \geq 0.$$

  Let $u = 1$, i.e. $\mathbb{E}_h[X] = 1$.

  All importance sampling simulations are executed with sample size 1000. The experiments show that the SCV is convex unimodal as a function of $\rho > 1$, that is, the SCV decreases until some ρ_0 and increases thereafter.

  1. $n = 1, \gamma = 20$.
     Probability: 2.06e-009
     SCV κ_g^2 of Sanov solution: 0.021954
     Table of SCV κ_f^2 of minxent-2 solution for increasing ρ:

ρ	5	10	50	100	200	500
κ_f^2	0.009929	0.007488	0.003189	0.002427	0.002009	0.001736

  2. $n = 5, \gamma = 30$.
     Probability estimate: 3.85e-009
     SCV κ_g^2 of Sanov solution: 0.012720
     Table of SCV κ_f^2 of minxent-2 solution for increasing ρ:

ρ	10	25	50	100
κ_f^2	0.004503	0.002955	0.002498	0.002269

  3. $n = 10, \gamma = 40$.
     Probability estimate: 4.11e-009
     SCV κ_g^2 of Sanov solution: 0.011061
     Table of SCV κ_f^2 of minxent-2 solution for increasing ρ:

ρ	10	25	50	100
κ_f^2	0.003500	0.002791	0.002392	0.002059

7 Conclusion

We have shown in a few simple static examples that the minxent approach for finding importance sampling densities is promising and leads to better results than the traditional exponentially tilted solution. Current research focusses on the application to heavy tailed distributions and to dynamic queueing models.

References

[1] Asmussen S. and R. Rubinstein, Steady-state rare events simulation in queueing models and its complexity prperties, *Advances in Queueing: Models, Methods and Problems* (ed. J. Dshalalow), 429 – 466, CRC Press, 1995.

[2] Cover T.M. and Thomas J.A., *Elements of Information Theory*, John Wiley & Sons, inc, 1991.

[3] Dembo A. and O. Zeitouni, *Large deviations techniques and applications*, Second edition. Springer Verlag, 1996.

[4] Kapur J.N. and H.K. Kesavan, *Entropy Optimization with Applications*, Academic Press, Inc., 1992.

[5] Rubinstein R.Y. and Kroese D.P., *The Cross-Entropy Method: a Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning*, Springer, 2004.