Minimum Cross-Entropy Methods for Rare-Event Simulation

Ad Ridder (*Vrije University, Amsterdam*)

Reuven Rubinstein (*Technion, Haifa*), Thomas Taimre (*University of Queensland, Brisbane*)

RESIM; Trondheim; June 2012

Introduction

- Suppose that $\ell_n = \mathbb{P}(\mathbf{X}_n \in A_n)$ is a rare-event probability;
- $\ell_n \to 0$ as $n \to \infty$;
- Randomness in the system is given by $\mathbf{X}_n$ which might be a variable, vector, process, ...
- And suppose that ℓ_n is difficult to compute analytically or numerically;
- But easy to estimate by Monte Carlo simulation.

Monte Carlo Simulations

- f_n is PDF (prob. density) of $\mathbf{X}_n$;
- Standard:
  1. generate IID $\mathbf{X}_n(k)$'s using f_n;
  2. take sample mean of $\mathbb{1}\{\mathbf{X}_n(k) \in A_n\}$;
- Importance sampling: find PDF g_n;
  1. generate IID $\mathbf{X}_n(k)$'s using g_n;
  2. compute likelihood ratios $L_n(\mathbf{X}_n(k)) = f(\mathbf{X}_n(k))/g_n(\mathbf{X}_n(k))$;
  3. take sample mean of $Z_n(k) = L_n(\mathbf{X}_n(k))\mathbb{1}\{\mathbf{X}_n(k) \in A_n\}$;

This Talk

We shall investigate importance sampling algorithms that use solutions to auxiliary convex optimization programs (minimum cross-entropy method).

We shall investigate classical simple rare-event problems concerning random walks, and queues, construct simulation algorithms, and analyse the efficiency of the estimators.

Efficiency

Let Z_n be the unbiased estimator of ℓ_n associated with a simulation algorithm.

Definition

(i) Z_n is strongly efficient (has BRE) if

$$\limsup_{n\to\infty} \frac{\mathbb{V}ar[Z_n]}{\ell_n^2} < \infty.$$

(ii) Z_n is logarithmically efficient (or AO) if

$$\liminf_{n\to\infty} \frac{|\log \mathbb{V}ar[Z_n]|}{|\log \ell_n^2|} \geq 1.$$

For unbiased estimators you may replace $\mathbb{V}ar[Z_n]$ by $\mathbb{E}[Z_n^2]$.

Zero-Variance

$$g_n(\mathbf{x}) = \frac{f_n(\mathbf{x})\mathbb{1}\{\mathbf{x} \in A_n\}}{\ell_n}$$

gives

$$\mathbb{V}ar[Z_n] = 0.$$

Agenda

A. The essence of minimum cross-entropy;

B. Random walk in one dimension; single rare event;

C. Random walk in one dimension; multiple rare events;

D. Face-homogeneous RW (queueing model) in one dimensions; single rare event;

E. Random walk in two dimensions; non-convex rare event;

F. Face-homogeneous RW (queueing model) in two dimensions; non-convex rare event;

A. The Essence of Minimum Cross-Entropy (MCE)

- Let $\mathbf{X}_n$ be n-dimensional vector; with PDF f_n;
- Let $H_1, \ldots, H_m : \mathbb{R}^n \to \mathbb{R}$; $b_1, \ldots, b_m \in \mathbb{R}$;
- Define set $A_n = \bigcap_{i=1}^m \{\mathbf{x} \in \mathbb{R}^n : H_i(\mathbf{x}) \geq b_i\}$;
- Polytope when H_i are linear;
- Problem to compute $\ell_n = \mathbb{P}(\mathbf{X}_n \in A_n)$.

MCE Program

This is just one of many possibilities.

$$\inf_{g_n \geq 0} \left\{ \mathcal{D}_{\mathrm{KL}}(g_n | f_n) \ : \ \int g_n(\mathbf{x})\, d\mathbf{x} = 1, \ \int H_i(\mathbf{x}) g_n(\mathbf{x})\, d\mathbf{x} \geq b_i, i = 1, \ldots, m \right\}.$$

Note solution g_n is PDF for which $\mathbb{E}_{g_n}[H_i(\mathbf{X}_n)] \geq b_i$ for all contraints.

Solution

$$g_n(\mathbf{x}) = \frac{f_n(\mathbf{x}) \exp(\sum_{i=1}^m \lambda_i H_i(\mathbf{x}))}{\mathbb{E}_{f_n}[\exp(\sum_{i=1}^m \lambda_i H_i(\mathbf{X}_n))]};$$

where $\lambda_1, \ldots, \lambda_m \geq 0$ obtained by dual program

$$\max_{\lambda_i} \sum_{i=1}^m \lambda_i b_i + \mathbb{E}_{f_n}\left[\exp\left(\sum_{i=1}^m \lambda_i H_i(\mathbf{X}_n) \right) \right].$$

Restrictions

- Single constraint $\ell_n = \mathbb{P}(H(\mathbf{X}_n) \geq b)$;
- H is separable $H(\mathbf{x}) = \sum_{j=1}^{n} H_j(x_j)$;
- Independent jumps $f(\mathbf{x}) = \prod_{j=1}^{n} f_j(x_j)$.

MCE Solution

$$g(\mathbf{x}) = \prod_{j=1}^{n} \frac{f_j(x_j)\exp(\lambda H_j(x_j))}{\mathbb{E}_{f_j}[\exp(\lambda H_j(X_j))]};$$

where $\lambda = 0$ if $H(\mathbf{x}) \geq b$, else

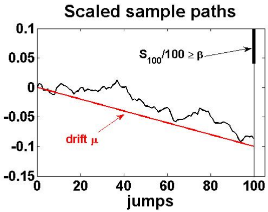
$$\frac{d}{d\lambda} \sum_{j=1}^{n} \log \mathbb{E}_{f_j}[\exp(\lambda H_j(X_j))] = b.$$

B. One-Tailed Problem For RW

Random walk (S_k) with IID jumps (X_j):

$$S_0 = 0$$

$$S_k = S_{k-1} + X_k = \sum_{j=1}^{k} X_j$$



Average of 10 sample paths for $n = 100$.

The Rare Event Problem

- Jump X has PDF $f(x)$.
- $f(x)$ has light positive and negative tails, i.e. $\mathbb{E}[\exp(\theta X)] < \infty$ for all θ in an open interval $(-\epsilon, \epsilon)$ containing zero.
- Problem: compute $\ell_n = \mathbb{P}(S_n \geq n\beta)$ for large n and $\beta > \mathbb{E}[X] = \mu$.

State-Dependent Importance Sampling

- $\mathbf{X}_n = (X_1, \ldots, X_n)$ vector of IID jumps;
- Has joint PDF $f_n(\mathbf{x}) = \prod_{j=1}^{n} f(x_j)$;
- IS using joint PDF $g_n(\mathbf{x})$;
- Allow to be not independent; not identical;
- (S_k) is allowed to become a time-inhomogeneous Markov chain with state-dependent jumps;
- Let $g_{k+1}(x|s)dx = \mathbb{P}(X_{k+1} \in (x, x+dx) \mid S_k = s)$;
- Thus (denoting $s_k = x_1 + \cdots + x_k$),

$$g_n(x_1, \ldots, x_n) = \prod_{k=0}^{n-1} g_{k+1}(x_{k+1}|x_1, \ldots, x_k) = \prod_{k=0}^{n-1} g_{k+1}(x_{k+1}|s_k).$$

MCE for Problem B

(R., Taimre 11)

- Notice that $g_{k+1}(x|s)$ is the marginal pdf of a joint pdf $g_{k+1\to n}(x_{k+1},\ldots,x_n|s)$ of all 'future' jumps $X_{k+1},\ldots,X_n$ given $S_k = s$.
- $g_{k+1\to n}(\cdot|s)$ is found as the solution of MCE program with single inequality constraint (MCE-IN):

$$\inf_{g\geq 0} \mathcal{D}_{\mathrm{KL}}(g|f_{k+1\to n})$$

$$\text{s.t.} \int g(\mathbf{x})\,d\mathbf{x} = 1,\ \mathbb{E}_g\Big[\sum_{j=k+1}^{n} X_j\Big] \geq n\beta - s.$$

Solution to MCE-IN

- Define $\psi(\theta) = \log \mathbb{E}[\exp(\theta X)]$ log MGF of jump;
- Set $\mu_k(s) = (n\beta - s)/(n-k)$ average remaining target drift;
- Then

$$g_{k+1}(x|s) = f(x)e^{\lambda x - \psi(\lambda)}$$

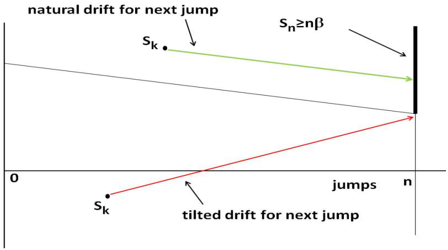
where λ satisfies

$$\begin{cases} \psi'(\lambda) = \mu_k(s), & \text{if } \mu_k(s) \geq \mu, \\ \lambda = 0, & \text{otherwise.} \end{cases}$$

Observations

- λ depends on time (k) and space (s);
- No biasing if the target set can be reached on average under the original distribution;
- Otherwise: the next jump is drawn from a distribution so that on average the process will drift to $n\beta$.

The possibility to 'turn off' the tilting.



Theorem

The MCE importance sampling estimator is logarithmically efficient.

- The proof is based on the property that the Markov chain (S_k) with the IS conditional jump densities $g_{k+1}(X_{k+1}|S_k)$ has the same fluid limit as under the state-independent tilting;
- Optimal state-independent tilting gives logarithmical efficiency.

Zero-Variance Approximation (ZVA)

(Blanchet & Glynn 06; L'Ecuyer, Blanchet, Tuffin & Glynn 10):

- The zero-variance IS pdf is

$$g_n^*(\mathbf{x}) = \frac{f_n(\mathbf{x})\mathbb{1}\{x_1 + \cdots + x_n \geq n\beta\}}{\ell_n}.$$

- Use jump PDF's of the form

$$g_k(x|s)dx = \mathbb{P}(X_k \in (x, x+dx) \,|\, S_{k-1} = s) = \frac{f(x)v_k(s+x)}{w_k(s)}dx,$$

  where $v_k(s)$ approximates $\mathbb{P}(S_n \geq n\beta | S_k = s)$.

- Apply also refinement of drawing the last jump from the original PDF f conditioned that $X_n \geq n\beta - S_{n-1}$. This makes the rare event certain to occur.

Theorem

The ZVA importance sampling estimator (with the refinement) has bounded relative error (BRE) in case of Gaussian jumps.

Proof given in (B&G06).

Efficiencies of IS Estimators (cont'd)

vrije Universiteit amsterdam

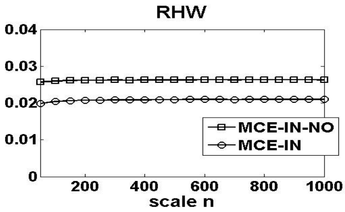
Theorem

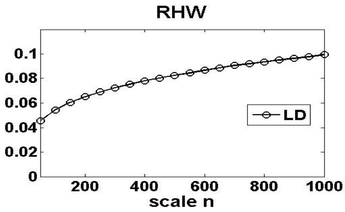
The MCE importance sampling estimator (with the refinement) has BRE for Gaussian jumps.

The proof is based on the similarity between the MCE and ZVA estimators. And some careful algebraic manipulations.

Simulation Results for $N(0, 1)$ Jumps

Relative half width (RHW) is 1.96 times the relative error.
Overflow level $\beta = 2/3$. Sample size $m = 10000$. Scaling n ranges 50-1000.
Results are averages of 100 repetitions.

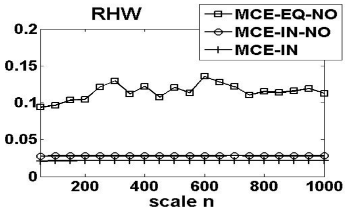


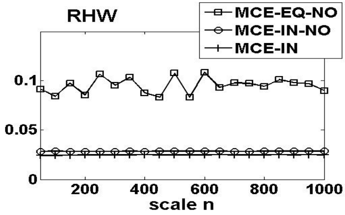


Other Jump Distributions

vrije Universiteit amsterdam

- Left: X has Laplace distribution ('double exponential') with mean 0 and variance 2. Overflow level $\beta = 1$.
- Right: X has double Coxian distribution with mean 0 and variance 6. Overflow level $\beta = 1.5$.





What for Heavy Tails?

(Asmussen plus co-authors)

- IID subexponential X_j's with a concave hazard rate function $\Lambda(x) = -\log \mathbb{P}(X_j > x)$;
- Examples: Weibull (with shape parameter < 1), Pareto (finite mean), Lognormal;
- Note only positive jumps;
- Goal $\ell_n = \mathbb{P}(\sum_{j=1}^{N} X_j > n)$ for fixed small N and $n \to \infty$.

(R., Rubinstein 07)

Recall $f_N(\mathbf{x})$ original joint PDF (product); g_N target IS joint PDF.

1. $\inf_{g_N \geq 0} \ \mathcal{D}_{\mathrm{KL}}(g_N | f_N)$s.t. $\int g_N(\mathbf{x})\, d\mathbf{x} = 1, \ \mathbb{E}_{g_N}\left[\sum_{j=1}^{N} \Lambda(X_j) \right] \geq \Lambda(n)$.
2. Constraint $\mathbb{E}_{g_N}\left[\Lambda\left(\sum_{j=1}^{N} X_j \right) \right] \geq \rho \Lambda(n)$ ($\quad 0 < \rho < 1$).

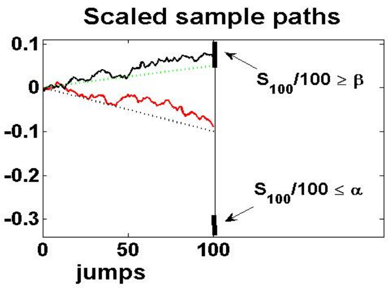
Solutions

- Program 1 gives (independent) hazard rate twisting (Juneja & Shahabuddin 02);
- Program 2 gives (correlated) hazard rate twisting;
- Both logarithmically efficient;
- Program 2 superior, statistically;
- Drawback: generate samples;
- Work in progress on adapted versions.

C. Two-Tailed Problem for RW

vrije Universiteit amsterdam

- Compute: $\ell_n = \mathbb{P}(S_n \leq n\alpha \text{ or } S_n \geq n\beta)$.
- For the same random walk as above, where $\alpha < \mathbb{E}[X] < \beta$ such that
- $I(\alpha) > I(\beta)$ for the large deviations rate function $I(\cdot)$.
- Well-known counter example to the use of a single state-independent importance sampling scheme (Glassermann & Wang 97, Bucklew 04).



Mixed Estimators

Note: the rare event consists of two disjoint events

$$A_n(1) = \{S_n \leq n\alpha\} \text{ and } A_n(2) = \{S_n \geq n\beta\}.$$

Suppose that for $j = 1, 2$:

(i) there is an IS estimator of $\mathbb{P}(A_n(j))$;

(ii) it applies pdf $g_j(\boldsymbol{x})$ for the jumps $X_1, \ldots, X_n$;

(iii) it has associated likelihood ratio $L_j(\boldsymbol{x}) = f(\boldsymbol{x})/g_j(\boldsymbol{x})$;

(iv) corresponding (single-sample) estimator is given by

$$Z_n(j) = L_j(\boldsymbol{X})\mathbb{1}\{A_n(j)\}.$$

Implementation

For any n, let Δ_n be a rv on $\{1,2\}$ with positive probabilities $\pi_n(j)$.
Such that Δ_n is independent of the $Z_n(j)$'s.

The mixed estimator is defined by

$$\begin{aligned}
Z_n &= \sum_j \frac{1}{\pi_n(j)} \mathbb{1}\{\Delta_n = j\}\, Z_n(j) \\
&= \sum_j \frac{1}{\pi_n(j)} \mathbb{1}\{\Delta_n = j\}\, L_j(\boldsymbol{X}) \mathbb{1}\{A_n(j)\} \\
&= \sum_j \mathbb{1}\{\Delta_n = j\}\, \frac{f(\boldsymbol{X})}{\pi_n(j) g_j(\boldsymbol{X})} \mathbb{1}\{A_n(j)\}.
\end{aligned}$$

Estimator is unbiased.

Proposition

Theorem

Assume that there are finite constants c_j s.t.

$$\limsup_{n\to\infty} \frac{\mathbb{E}[Z_n^2(j)]}{(\mathbb{E}[Z_n(j)])^2} \leq c_j$$

for all j, i.e., all $Z_n(j)$ are strongy efficient. Then the mixed estimator is strongly efficient.

Proof follows by working out the second moment of the mixed estimator $\mathbb{E}[Z_n^2]$.

Proposition

Theorem

Assume for any j:

(a) *the sequence of probabilities $(\mathbb{P}(A_n(j)))_n$ satisfies a large deviations limit $\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(A_n(j)) = -I(j)$;*

(b) *the estimator $Z_n(j)$ is logarithmically efficient;*

(c) *the sequence of mixing probabilities $(\pi_n(j))_n$ may not tend to zero exponentially fast: $\lim_{n\to\infty} \frac{1}{n} \log \frac{1}{\pi_n(j)} = 0$.*

Then, the mixed estimator Z_n is logarithmically efficient.

Proof

1. Establish LD for $\mathbb{E}[Z_n]$ by applying the principle of the largest term:

$$\lim_{n\to\infty} \frac{1}{n} \log\mathbb{E}[Z_n] = \lim_{n\to\infty} \frac{1}{n} \log \sum_j \mathbb{E}\left[Z_n(j)\right] = \max_j \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}\left[Z_n(j)\right]$$
$$= -\min_j I(j) = -I.$$

2. Establish lower bound LD for $\mathbb{E}[Z_n^2]$ by applying Jensen's inequality:

$$\liminf_{n\to\infty} \frac{1}{n} \log\mathbb{E}\left[Z_n^2\right] \geq \liminf_{n\to\infty} \frac{1}{n} \log(\mathbb{E}[Z_n])^2$$
$$= 2 \liminf_{n\to\infty} \frac{1}{n} \log \mathbb{E}[Z_n] = -2I.$$

3. Establish upper bound LD for $\mathbb{E}[Z_n^2]$ (more involved algebra).

Random or Deterministic Mixing?

- Sample size m gives mixed 'randomized' estimator:

$$Z_n^{\mathrm{r}}[m] = \frac{1}{m} \sum_{i=1}^{m} \sum_{j} \frac{1}{\pi_n(j)} \mathbb{1}\{\Delta_n^{(i)} = j\} Z_n^{(i)}(j).$$

- Mixing deterministic fractions gives mixed 'deterministic' estimator:

$$Z_n^{\mathrm{d}}[m] = \sum_{j} \frac{1}{m_n(j)} \sum_{i=1}^{m_n(j)} Z_n^{(i)}(j),$$

where $m_n(j) = [\pi_n(j) m]$.

Clearly $\mathbb{V}ar\left[Z_n^{\mathrm{d}}[m]\right] \leq \mathbb{V}ar\left[Z_n^{\mathrm{r}}[m]\right]$.

The Mixing Probabilities

- Minimise asymptotically the variance of the 'deterministic' estimator using the large deviations rate expressions [Glasserman&Wang 97]:

$$\pi_n(1) = \frac{\exp\big(-nI(a)+o(n)\big)}{\exp\big(-nI(a)+o(n)\big)+\exp\big(-nI(b)+o(n)\big)}$$
$$= \frac{1}{1+\exp\big(n(I(a)-I(b))+o(n)\big)} \approx \exp\big(-n(I(a)-I(b))\big).$$

  Cut-off to prevent exponential decaying to zero: $\pi_n(1)\vee\eta$.

- Apply again MCE: gives numerical values for any n, asymptotically as $n\to\infty$ similar as above.

1. MCE: mixing the MCE-IN estimators of the single-tail problems.
2. LD: mixing the LD estimators.
3. DW-SOL: state-dependent algorithm of [Dupuis&Wang 04].

   In words, this algorithm is doing the following. At any time it detects which of the two parts of the rare event is the most likely one, and then applies an exponential tilting of the next jump X_k in order to get there on average.

4. DW-SUBSOL: state-dependent algorithm of [Dupuis&Wang 07].

   Each jump is realised from a mixture of exponentially tilted densities, i.e.,

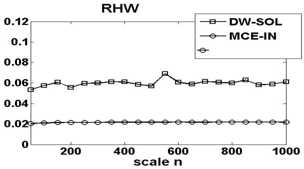
$$\mathbb{P}(X_{k+1} \in (x, x+dx) | S_k = s) = \sum_{j=1}^{2} \pi_j^\delta f(x) \exp(\theta_j x - \psi(\theta_j)) dx.$$

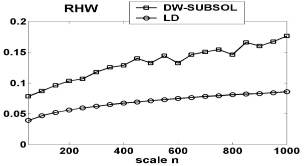
   The tilting parameters θ_j are fixed throughout the simulation, and the mixing probabilities π_j^δ depend on jump time $k+1$, state $S_k = s$, and so-called mollification parameter δ.

   Parameters follow from solving a so-called Isaacs equation.

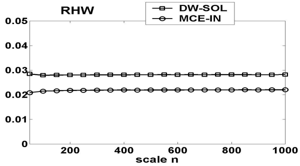
Simulation Results

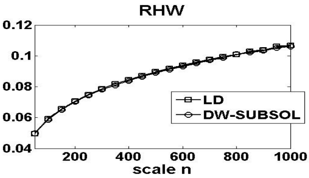
Gaussian jumps: $\mu = 3, \sigma^2 = 1$. Overflow level $\alpha = 2.4999; \beta = 3.5$.
Sample size $m = 10000$. Scaling n ranges 50-1000.
Results are averages of 100 repetitions.





Laplacian jumps: $\kappa = 1$. Overflow level $\alpha = -1.25; \beta = 1.0$.





D. Face-Homogeneous One-Dimensional RW

Examples: $M^X/M/1$, $M/G/1$ (stable).

- (S_k) the state process (after embedding at jump times).
- For $\Lambda \in \{0,1\}$: $(X_k^{(\Lambda)})$ i.i.d. geometric jumps,

$$a_j = \mathbb{P}(X^{(0)} = j) = (1-p^{(0)})(p^{(0)})^j, \quad j = 0,1,\dots,$$
$$b_j = \mathbb{P}(X^{(1)} = j) = (1-p^{(1)})(p^{(1)})^{j+1}, \quad j = -1,0,\dots.$$

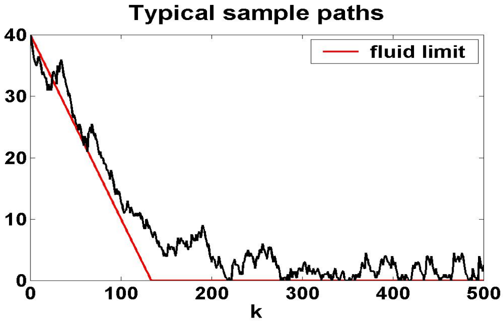
- Define $\Lambda(x) = \mathbb{1}\{x > 0\}$, then

$$S_k = S_{k-1} + X_k^{(\Lambda(S_{k-1}))}, \quad k = 1,2,\dots.$$

- Mean jumps:

$$\mu^{(0)} = \frac{p^{(0)}}{1-p^{(0)}} > 0, \quad \mu^{(1)} = -\frac{1-2p^{(1)}}{1-p^{(1)}} < 0.$$

Typical Sample Paths



Data: $\mu^{(0)} = 0.5, \mu^{(1)} = -0.3.$

Rare Event Problem

vrije Universiteit amsterdam

- Compute: $p_n = \mathbb{P}(S_{nT} \geq n\beta | S_0 = ns_0)$;
- or, equivalently after scaling $S^{[n]}(t) = S_{nt}/n$:

$$\mathbb{P}(S^{[n]}(T) \geq \beta | S^{[n]}(0) = s_0),$$

- for large n, and fixed given initial state $s_0 \geq 0$, target level $\beta > 0$ and horizon T.

Sample Path Large Deviations

[Shwartz-Weiss 1995].

- Concept of path. $y : [0, T] \to \mathbb{R}_+$ absolute continuous with $y(0) = s_0$.
- Most likely path $y^{[\infty]}$ given by

$$y^{[\infty]}(t) = \begin{cases} s_0 + \mu^{(1)} t & 0 \le t \le t_0, \text{ where } t_0 = -s_0/\mu^{(1)} \\ 0 & t_0 \le t \le T. \end{cases}$$

- Meaning $\lim_{n\to\infty} \mathbb{P}\left(\sup_{0\le t\le T} \left| S^{[n]}(t) - y^{[\infty]}(t) \right| < \epsilon\right) = 1$.
- Large deviations for piecewise linear paths (generalized to absolute continuous paths) $y : [0, T] \to \mathbb{R}_{\ge 0}$:

$$\lim_{\epsilon\downarrow 0} \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\Big(\sup_{0\le t\le T} \left| S^{[n]}(t) - y(t) \right| < \epsilon \Big) = -J(y).$$

- Optimal path to the rare event: $y^* = \arg\min\{J(y) : y(T) \ge \beta\}$.

There is a critical horizon T^*, such that

- for small horizon T $(T < T^*)$; optimal path is straight line to rare event:

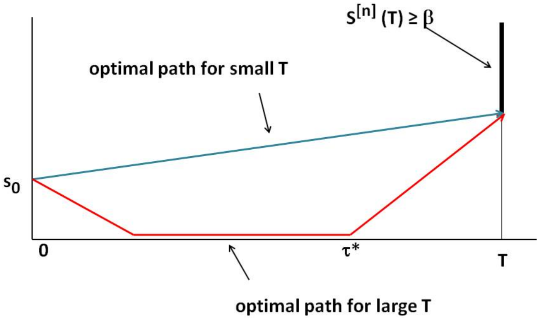
$$y^{(1)}(t) = s_0 + v^{(1)} t \quad \text{with 'speed'} \quad v^{(1)} = (\beta - s_0)/T$$

- Large horizon T $(T < T^*)$; optimal path follows the most likely path until a a switching time τ^*, and then goes straight to the rare event:

$$y^{(0)}(t) = \begin{cases} y^{[\infty]}(t), & 0 \le t \le \tau^* \\ v^{(0)}(t - \tau^*), & (\tau^* \le t \le T) \end{cases} \quad \text{with 'speed'} \quad v^{(0)} = \beta/(T - \tau^*).$$

Optimal Path to Overflow

vrije Universiteit amsterdam



Importance Sampling

- Bias the (distributions of the) jump variables $X^{(\Lambda)}$.
- They remain geometric, but with other success probability (other mean jump).
- Allow biasing to depend on time (k) and state ($S_k = s$).
- Denote the mean jump ('drift') at time k in scaled state s under the change of measure by

$$\mu_k^{\mathrm{IS}}(s) = \mathbb{E}^{\mathrm{IS}}[X_k | S_{k-1}/n = s].$$

Assumption

We consider the case of large horizon, $T > T^*$.

Algorithm D(i)

- State independent (constant biasing).
- $\mu_k^{\mathrm{IS}}(s) = (\beta - s_0)/T$.
- The scaled sample paths follow path $y^{(1)}$.

Algorithm D(ii)

- Time dependent biasing.
- No biasing until $k = n\tau^*$. Else

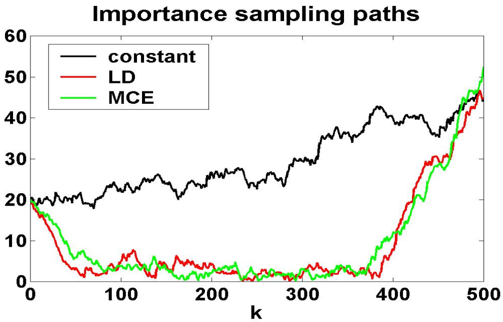
$$\mu_k^{\mathrm{IS}}(s) = \beta/(T - \tau^*).$$

- The scaled sample paths follow path $y^{(0)}$.

Algorithm D(iii)

- Recursively; state and time dependent.
- Based on the MCE algorithm for RW's: at any time $t = k/n < \tau^*$, $s = S_k/n$ is starting point of a scaled overflow problem with horizon $T - (k/n)$, called the reduced problem.
- Find critical horizon, and switching time of the reduced problem.
- Set $\mu_k^{\mathrm{IS}}(S_k/n)$ so that the scaled process will 'optimally drift' to β, i.e., follows the optimal path in the reduced problem.

Simulated Paths



Proposition

Theorem

Algorithms D(ii) and D(iii) are logarithmically efficient.

The proof is based on two issues:

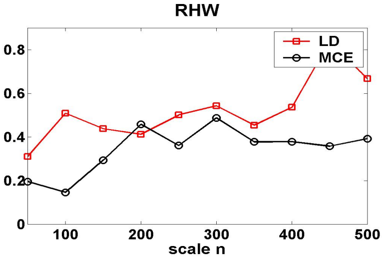
1. The sample path large deviations:

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}[Z_n] = \lim_{n\to\infty} \frac{1}{n} \log p_n = -J(y^*).$$

2. Show that the second moment of the estimator Z_n satisfies

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{E}[Z_n^2] \leq -2J(y^*).$$

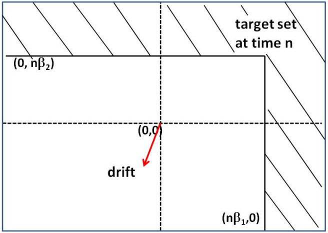
Parameters: $\mu^{(0)} = 0.5, \mu^{(1)} = -0.3, s_0 = 0.4, \beta = 1, T = 10 > T^*$.
Sample size $m = 10000$. Scaling n ranges 10-500. Results are averages of 10 repetitions.



E. Two-Dimensional RW

Jump $X = (X_1, X_2) \in \mathbb{R}^2$, with PDF $f(x)$. $S_0 = 0 = (0,0)$. $S_k = S_{k-1} + X_k$ where $X_1, X_2, \ldots$ i.i.d. (abusing notation).

Negative drift (both coordinates) $\mu = (\mu_1, \mu_2) = \mathbb{E}[X] < 0$.



Problem

Compute (or simulate):

$$p_n = \mathbb{P}(S_{n,1} \geq n\beta_1 \text{ or } S_{n,2} \geq n\beta_2 \,|\, S_0 = 0),$$

for fixed scaled target level $\beta = (\beta_1, \beta_2) \in \mathbb{R}^2_+$, and parameter $n \to \infty$.

Objective

- Define $A_n(j) = \{S_{n,j} \geq n\beta_j\}$ for $j = 1, 2$.
- Goal: construct a logarithmically efficient estimator for $\mathbb{P}(A_n(1) \cup A_n(2))$.
- Difficulty: the rarity set is not convex (Dupuis&Wang 07).

Assumption

There are importance sampling estimators $Z_n(j)$ for $\mathbb{P}(A_n(j))$ that implement pdf $g_j(\boldsymbol{x})$ for the jumps $X_1, \ldots, X_n$.

Thus for $j = 1, 2$:

$$Z_n(j) = \frac{f(\boldsymbol{x})}{g_j(\boldsymbol{x})} \mathbb{1}\{A_n(j)\}.$$

Mixed Estimator

Recall the mixed estimator that randomly mixes the $Z_n(j)$'s:

$$\begin{aligned} Z_n^{\mathrm{r}} &= \sum_j \frac{1}{\pi_n(j)} \mathbb{1}\{\Delta_n = j\}\, Z_n(j) \\ &= \sum_j \mathbb{1}\{\Delta_n = j\}\, \frac{f(\boldsymbol{X})}{\pi_n(j) g_j(\boldsymbol{X})} \mathbb{1}\{A_n(j)\} \end{aligned}$$

Estimator is biased because events $A_n(j)$ are not disjunct.

Mixture Estimator

Definition

The mixture importance sampling estimator is defined by

$$Z_n^{\mathrm{mix}} = \frac{f(\boldsymbol{X})}{\sum_j \pi_n(j) g_j(\boldsymbol{X})} \mathbb{1}\{\bigcup_j A_n(j)\}.$$

Estimator is unbiased.

Proposition

Theorem

Assume that all $Z_n(j)$ are strongly efficient, i.e., there are finite constants c_j s.t.

$$\limsup_{n\to\infty} \frac{\mathbb{E}[Z_n^2(j)]}{(\mathbb{E}[Z_n(j)])^2} \leq c_j.$$

Then both the mixed estimator Z_n^{r} and the mixture IS estimator Z_n^{mix} are strongly efficient.

Proof follows by working out the second moment of the estimators.

Proposition

Theorem

Assume for any j:

(a) *the sequence of probabilities $(\mathbb{P}(A_n(j)))_n$ satisfies a large deviations limit $\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(A_n(j)) = -I(j)$;*

(b) *the estimator $Z_n(j)$ is logarithmically efficient;*

(c) *the sequence of mixing probabilities $(\pi_n(j))_n$ may not tend to zero exponentially fast: $\lim_{n\to\infty} \frac{1}{n} \log \frac{1}{\pi_n(j)} = 0$.*

Then, both the mixed estimator Z_n^{r} and the mixture IS estimator Z_n^{mix} are logarithmically efficient.

Proof

Follows the same reasoning as above (model C); and

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(A_n(1) \cup A_n(2))$$
$$= \lim_{n\to\infty} \frac{1}{n} \log \left(\mathbb{P}(A_n(1)) + \mathbb{P}(A_n(2))\right) = -\min_j I(j),$$

and

$$\mathbb{E}[(Z_n^{\text{mix}})^2] \leq \mathbb{E}[(Z_n^{\text{r}})^2].$$

Problem

Find the importance sampling densities $g_j(\boldsymbol{x})$ so that the components $Z_n(j)$ are logarithmically efficient.

1. Retain i.i.d. jumps by state-independent exponential tilting:

$$g_j(x_1, x_2) \propto f(x_1, x_2) e^{\theta_1(j) x_1 + \theta_2(j) x_2}.$$

   The tilting factor θ_1:

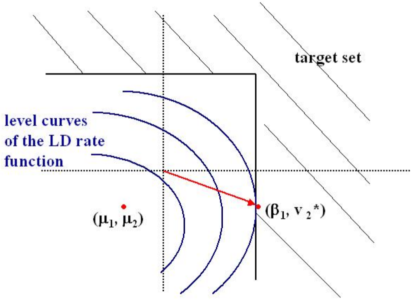
$$\min_{v_2 \in \mathbb{R}} \sup_{\theta} \left(\theta_1 \beta_1 + \theta_2 v_2 - \psi(\theta_1, \theta_2)\right).$$

   Similar for θ_2.

2. Apply MCE to construct state-dependent importance sampling densities $g_j(\boldsymbol{x})$ for the two components.

No results yet: work in progress.

In a Figure



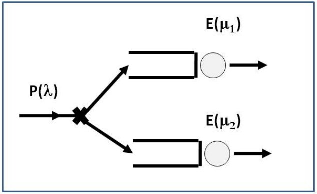
The red arrow is the drift of the jumps in the component to the right.

F. Fork-Join Queue

Special case of two-dimensional face-homogeneous random walk.

Model

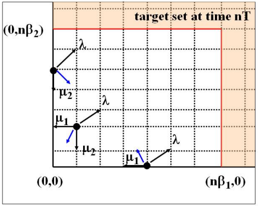
- Poisson (λ) arrivals;
- an arriving job splits in two subjobs;
- two independent single server queues;
- exponential service times with rate μ_1 and μ_2, resp;
- for stability $\lambda < \min(\mu_1, \mu_2)$.



[Flatto & Hahn 84]

The Rare Event

$\{S_k = (S_{k,1}, S_{k,2}) : k = 0, 1, \ldots\}$ is the discrete-time Markov chain analogon of the fork-join queue by embedding at jump times;
S_k represents the backlogs at the queues.



Problem

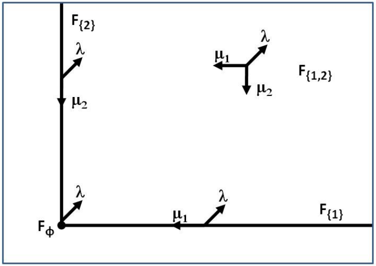
Estimate by simulation:

$$p_n = \mathbb{P}(S_{1,nT} \geq n\beta_1 \text{ or } S_{2,nT} \geq n\beta_2 | S_0 = ns_0),$$

for fixed scaled initial state $s_0 = (s_{0,1}, s_{0,2}) \in \mathbb{R}_+^2$, fixed scaled threshold $\beta = (\beta_1, \beta_2) \in \mathbb{R}_+^2$, fixed scaled horizon $T > 0$, and parameter $n \to \infty$.

Face-Homogeneous RW

The fork-join queue is a face-homogeneous random walk on $\mathbb{Z}_+^2$ with four faces.



- The transition probabilities $p_{s,s+d}$ are constant for s in the same face $F^{(\Lambda)}$.
- On each face we might associate a random walk $(S_k^{(\Lambda)})_{k=0}^{\infty}$ with jump variable $X^{(\Lambda)}$ with probabilities $p^{(\Lambda)}(j) \doteq p_{s,s+j}$.

Importance Sampling Scheme

- Our importance sampling scheme will be a mixture of two sets of exponentially shifted jump probabilities of the jump variables $X^{(\Lambda)}$.
- For any $\theta \in \mathbb{R}^2$, the θ-shifted jump $X_\theta^{(\Lambda)}$ has jump probabilities

$$p_\theta^{(\Lambda)}(j) = e^{\langle \theta, j \rangle - \psi^{(\Lambda)}(\theta)} p^{(\Lambda)}(j),$$

  where $\psi^{(\Lambda)}(\cdot)$ is the log moment generating function of jump variable $X^{(\Lambda)}$.
- This gives us a set of 4 jump (or transition) probability densities.
- We have two of such sets, and before we simulate a sample path, we choose randomly a set.

Furthermore

Assumption: the time horizon is large.

- An optimal importance sampling is time-dependent in the sense, that
- after having chosen a set of biased jump probabilities, the biasing starts not before a switching time τ^* (cf. the one-dimensional problem).

No Details

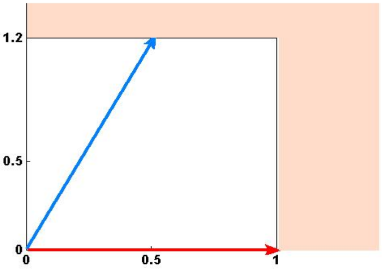
The details follow by applying

- sample path large deviations for phase-homogeneous random walks (Ignatiouk 01, 05);
- universal simulation distributions (Bucklew et al. 90, 04);
- numerical calculations for the optimal shift factors θ^* and the optimal switching times τ^*.

Under certain conditions, the mixture importance sampling scheme is logarithmically efficient (Bucklew et al. 90, 04)

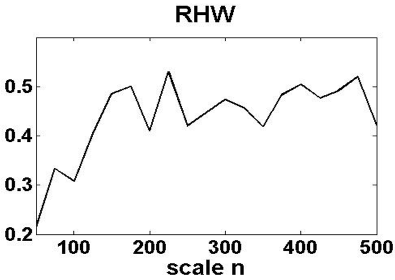
Example

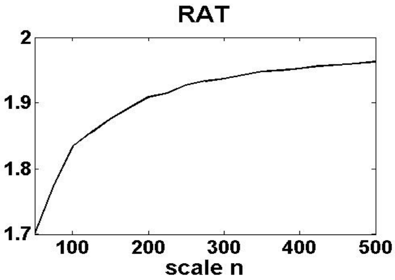
$\lambda = 1, \mu_1 = 1.5, \mu_2 = 2, s_0 = (0,0), \beta = (1, 1.2), T = 10.$
The two sets of biasing schemes are mixed by 0.8 (red path) and 0.2 (blue path).



Example (cont'd)

Results for scalings $n = 25–500$ with sample size $k = 50000$.





Conclusion and Outlook

- MCE seems a promising method for obtaining efficient importance sampling estimators in random walk environments.
- Higher dimensional rare-event queueing problems are nontrivial and need more advanced techniques, such as mixtures of important sampling algorithms.
- Further invastigations include time and state dependent MCE for these queueing problems.