The cross-entropy method for importance sampling simulation of the infinite-server queue

Ad Ridder

Department of Econometrics
Vrije Universiteit Amsterdam

INFORMS Applied Probability, Eindhoven, 2007

Outline

Outline

1. Model and problem
2. Large deviations
3. Importance sampling
4. Heuristics
5. Numerical results
6. Conclusion

The $M/G/\infty$ queueing model

- Poisson λ arrivals.
- General service time with cdf F and mean $1/\mu$.
- Infinitely many servers: upon arrival service starts immediately.
- $X(t)$ is number of busy servers at time t $(t \geq 0)$.
- $X(0) = 0$.

The level crossing problem

- First passage times:

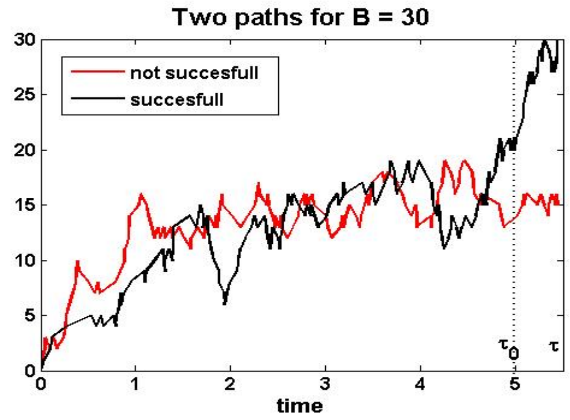
$$T(\ell) := \inf\{t \geq 0 : X(t) \geq \ell\}, \; \ell = 1, 2, \ldots .$$

- Problem: given level B and times τ_0, τ $(0 \leq \tau_0 < \tau)$ find

$$P\Big(T(B) \in (\tau_0, \tau]\Big).$$

- Assumptions: B is large and $\lambda/\mu < B$.
- $t \to \infty$ gives the stationary regime where $X(\infty)$ is Poisson with mean λ/μ.

A plot of two realisations



The n-systems

- Let $\lambda = \lambda_n$ and $B = B_n$ $(n = 1, 2, \ldots)$ grow proportionally to n according to

$$\lambda_n = n\gamma, \quad B_n = nb,$$

where γ and b fixed, and satisfy $\gamma/\mu < b$.
- We have for each n an infinite server system.
- $X_n(t)$ are the occupancies, $T_n(\ell)$ the first passage times in the n-system.
- The probability becomes

$$p_n := P\Big(T_n(nb) \in (\tau_0, \tau]\Big).$$

- We set (w.l.o.g.) $b = 1$

Outline

Large deviations

Theorem

$$-\lim_{n\to\infty} \tfrac{1}{n} \log p_n = \rho(\tau) - b\log\rho(\tau) + b\log b - b,$$

where $\rho(t) = \gamma \int_0^t (1 - F(x))\, dx.$

Proof

Step 1. Well-known that for any $t > 0$ (recall $X_n(0) = 0$)

$X_n(t) \stackrel{\mathrm{d}}{=} \sum_{i=1}^n X^{(i)}(t)$, where

$X^{(1)}(t), \ldots, X^{(n)}(t)$ are i.i.d. with Poisson-$\rho(t)$ distribution.

LD proof (cont'd)

Step 2. Apply Cramér's Theorem:

$$\lim_{n\to\infty} \tfrac{1}{n} \log P(X_n(t) \geq nb) = \lim_{n\to\infty} \tfrac{1}{n} \log P\left(\frac{1}{n}\sum_{i=1}^n X^{(i)}(t) \geq b\right) = -I_t(b),$$

where the large deviations rate function

$$I_t(b) = \sup_\theta \left(\theta b - \psi_t(\theta)\right),$$

with logarithmic moment generating function

$$\psi_t(\theta) = \log E\left[\exp(\theta X^{(\cdot)}(t))\right].$$

Doing the calculus gives $I_t(b)$ the expression of the Theorem.

LD proof (cont'd)

Step 3. Define

$$A_n = \bigcup_{t\leq\tau_0} \{X_n(t) \geq nb\}, \quad B_n = \bigcup_{\tau_0<t\leq\tau} \{X_n(t) \geq nb\}.$$

Thus, $p_n = P(A_n^c \cap B_n)$.
Upper bound:

$$\limsup_{n\to\infty} \tfrac{1}{n} \log p_n = \limsup_{n\to\infty} \tfrac{1}{n} \log P(A_n^c \cap B_n)$$

$$\leq \limsup_{n\to\infty} \tfrac{1}{n} \log P(B_n) = -\inf_{\tau_0<t\leq\tau} I_t(b) = -I_\tau(b),$$

applying Laplace's principle and that $I_t(b)$ decreases (as a function of t).

LD proof (cont'd)

Step 4. Lower bound.

$$p_n = P(A_n^c \cap B_n) = P(B_n) - P(A_n \cap B_n)$$

$$\geq P(B_n) - P(A_n) = P(B_n)\left(1 - \frac{P(A_n)}{P(B_n)}\right).$$

And

$$\liminf_{n\to\infty} \tfrac{1}{n} \log p_n \geq \liminf_{n\to\infty} \tfrac{1}{n} \log P(B_n)\left(1 - \frac{P(A_n)}{P(B_n)}\right)$$

$$\geq \liminf_{n\to\infty} \tfrac{1}{n} \log P(B_n) + \liminf_{n\to\infty} \tfrac{1}{n} \log\left(1 - \frac{P(A_n)}{P(B_n)}\right)$$

$$\geq -I_\tau(b) + \liminf_{n\to\infty} \tfrac{1}{n} \log \frac{1}{2} = -I_\tau(b).$$

Outline

Importance sampling

- Simulation of the infinite server model for estimation the probability.
- Importance sampling because level crossing is a rare event.
- Estimator based on N runs

$$Y_n^* := \frac{1}{N} \sum_{i=1}^{N} L(\{X_n^{(i)}(t), 0 \leq t \leq \tau\}) \mathbb{1}\{T_n^{(i)}(nb) \in (\tau_0, \tau]\}.$$

Exponential servers

In the exponential model we can derive

Done previously for exponential servers

- a sample path large deviations;
- a most likely ('optimal') path to overflow;
- a continuous shift function $\theta^*(t) : [0, \tau] \to \mathbb{R}_{\geq 0}$ such that importance sampling with arrival rates $\lambda\, e^{\theta^*(t)}$ and service rates $\mu\, e^{-\theta^*(t)}$ is asymptotically optimal:

$$\lim_{n\to\infty} \frac{\log E[(Y_n^*)^2]}{\log p_n} = 2.$$

Algorithm updates all realised services (of present customers) after each jump (arrivals and departures).

Outline

General service times

No memoryless property: updating of all services is 'impossible'.

The importance sampling algorithm

- The interval $[0, \tau]$ is partitioned in K equal subintervals I_k.
- The arrival rate on I_k is $\lambda\, e^{\theta_k}$.
- The service distribution of arriving customers in I_k is an exponentially shifted version of the original F, with shift parameter δ_k.
- No updates of service times of the other customers already present; no updates at a departure epoch.

Exponentially shifted distribution

Service time S has cdf F with density f.

Shifting with parameter δ:

$$f^{\delta}(x) = \frac{e^{\delta x} f(x)}{M(\delta)},$$

where $M(\delta)$ normalizing constant (moment generating function).

Denote $\psi(\delta) = \log M(\delta)$.
The expectation of S with the shifted distribution:

$$E^{\delta}[S] = \psi'(\delta).$$

The importance sampling parameters

Problem : which importance sampling parameters $\theta = (\theta_k)_{k=1}^{K}$ for arrivals and $\delta = (\delta_k)_{k=1}^{K}$ for services?

Idea: use the parameters from the exponential model:

$$\theta_k = \theta^*(t_k), \quad \psi'(\delta_k) = e^{\theta^*(t_k)}/\mu,$$

where t_k is the midpoint of the k-th subinterval I_k.

And $\theta^*(t)$ is the continuous shift parameter in the exponential model which is available in a closed form expression.

Simulation results

Model: $\gamma = 0.5, E[S] = \mu^{-1} = 1, b = 1, \tau_0 = 5.0, \tau = 5.5$, and

Coxian service times with two phases, and squared coefficient of variation (SCV) 5:

$$S \stackrel{d}{=} \Delta \mathrm{Exp}(\mu_1) + (1 - \Delta)\left(\mathrm{Exp}(\mu_1) + \mathrm{Exp}(\mu_2)\right),$$

where Δ is Bernoulli(p).

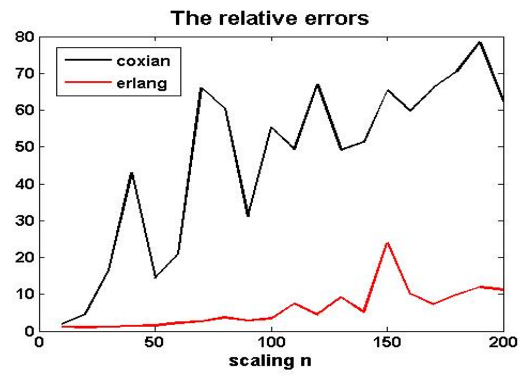
Erlangian service times with two phases, and SCV 0.5:

$$S \stackrel{d}{=} \mathrm{Exp}(2\mu) + \mathrm{Exp}(2\mu).$$

After exponential shifting Coxian remains Coxian and Erlang remains Erlang.

Plot

Scaling $n = 10, 20, \ldots, 200$. $K = 20$ subintervals.
IS-simulation: 50,000 runs.
Plot of the relative errors (in %).



Cross-entropy

We shall improve the Coxian case by applying the cross-entropy method for finding the shift parameters.

That is: solve

$$\max_{\theta,\delta} E\left[Y_n \log H\Big(\{X_n(t), 0 \le t \le \tau\} | \theta, \delta\Big)\right],$$

where $Y_n = 1\{T_n(nb) \in (\tau_0, \tau]\}$ indicates the occurrence of the rare event,

and $H(\cdot)$ the likelihood of the sample path when simulating according to the importance sampling algorithm with shift parameters θ and δ.

Solving the maximum likelihood

Because of the availability of an explicit expression for the likelihood, and by interchanging expectation and differentation, we can solve the first order conditions.
For $k = 1, \ldots, K$:

$$\frac{\partial}{\partial \theta_k} E[Y_n \log H(\cdot|\theta, \delta)] = 0 \;\Leftrightarrow\; \lambda\, e^{\theta_k} = \frac{E[Y_n N_k]}{E[Y_n \sum_{j=1}^{N_k} A_j]},$$

$$\frac{\partial}{\partial \delta_k} E[Y_n \log H(\cdot|\theta, \delta)] = 0 \;\Leftrightarrow\; \psi'(\delta_k) = \frac{E[Y_n \sum_{j=1}^{N_k} S_j]}{E[Y_n N_k]}.$$

Where N_k is the number of arrivals during subinterval I_k, with corresponding interarrival times A_j and service time S_j.

Cross-entropy algorithm

The expectations in the f.o.c. equations are estimated by simulation.

Since they involve the rare event (rv Y_n) we use importance sampling with θ and δ determined in the previous iteration.

Cross-entropy algorithm

1. Choose initial $\theta_k^{(0)}$ and $\delta_k^{(0)}$, $k = 1, \dots, K$; $i = 0$.
2. Simulate the infinite server queue $\{X_n(t) : 0 \le t \le \tau\}$ with arrival rates $\lambda \exp(\theta_k^{(i)})$ and shifted service time distributions with parameters $\delta_k^{(i)}$.
3. Estimate by importance sampling the expections $E[Y_n N_k]$, $E[Y_n \sum_{j=1}^{N_k} A_j]$, and $E[Y_n \sum_{j=1}^{N_k} S_j]$.
4. Find the updated $\theta_k^{(i+1)}$ and $\delta_k^{(i+1)}$.
5. Set $i = i + 1$ and repeat from 2 until convergence.

Outline

1. Model and problem
2. Large deviations
3. Importance sampling
4. Heuristics
5. Numerical results
6. Conclusion

Simulation

Same model with scaling $n = 50$.
$K = 20$ intervals; 20 CE-iterations of 5,000 samples.

Plots of initial parameters $\theta_k^{(0)}$, $\delta_k^{(0)}$ and after 20 iterations $\theta_k^{(20)}$, $\delta_k^{(20)}$ (as functions of k).

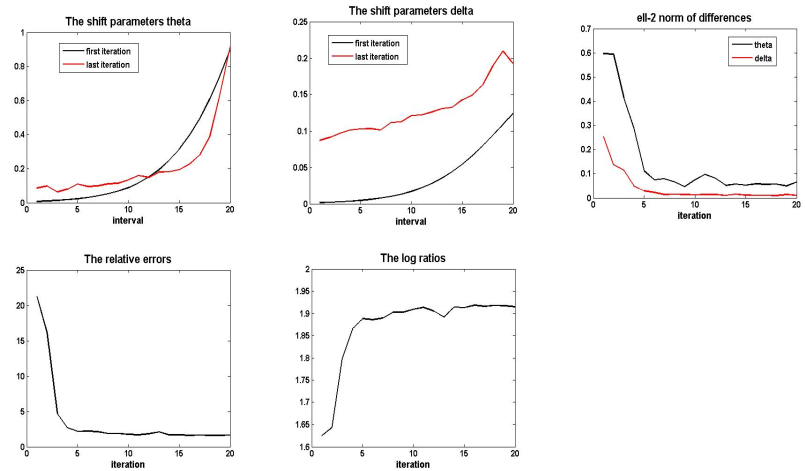
Plot of the 2-norms of the differences of two consecutiove solutions:

$$||\theta^{(i+1)} - \theta^{(i)}||_2, \quad ||\delta^{(i+1)} - \delta^{(i)}||_2.$$

After each CE-update we executed an IS simulation with 20,000 samples to estimate the rare-event probability p_n. Plot of the (estimated) relative errors and the (estimated) log ratios of the estimators:

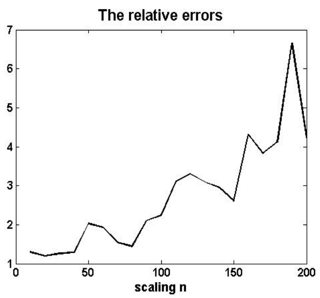
$$\mathrm{RE} = \frac{\sqrt{\mathrm{Var}[Y_n^*]}}{E[Y_n^*]}, \quad \mathrm{logratio} = \frac{\log E[(Y_n^*)^2]}{\log E[Y_n^*]}$$

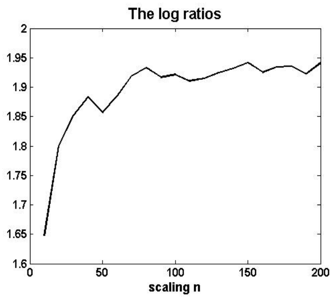
Plots for scaling $n = 50$



Larger scalings

Scaling $n = 10, 20, \ldots, 200$: $p_{200} \approx 3 \cdot 10^{-27}$.
CE-iterations: $\sim$ 10 to 20; 5000 runs each;
IS-simulation: 20,000 runs.
Plots of the relative errors (in %) and the log ratios.





How many CE-iterations?

Empirically: in the first iterations of th CE algorithm some of the θ_k and/or δ_k parameters become negative.

Most of the experiments gave all positive parameters within 10 iterations.

Good performance when all parameters became positive.

Implementation: stop CE updating after a few (for instance 5) iterations with all positive parameters.

Alternative CE algorithms

1. Start with initial parameters all equal to 0.
   That is: the original Monte Carlo simulation.

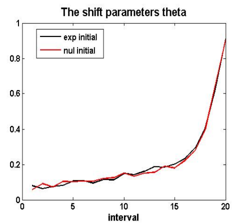
   Need to adapt the first few iterations to make sure that observations occur.
   Lower down the target level B. And increase it in each iteration based on the observations of the previous iteration.
2. Use smoothing in the updating rule:

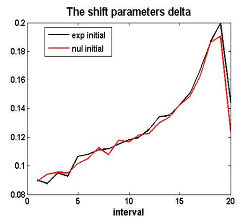
$$\delta_k^{(i+1)} = \alpha \widetilde{\delta_k}^{(i+1)} + (1-\alpha)\delta_k^{(i)},$$

where $\widetilde{\delta_k}^{(i+1)}$ follows the original updating.

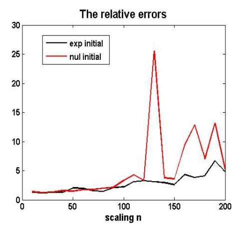
Results with the null initial

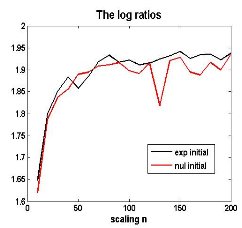
Plots of the shift parameters after 20 CE-iteations ($n = 50$).

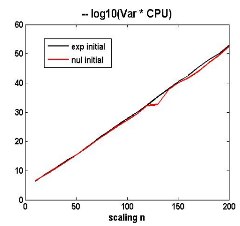




Plot of the relative errors, log ratios, and efforts for $n = 10, \ldots, 200$.







Heavy-tailed services

Experiments for Pareto with mean 1 and infinite variance:

$$f(x) = \frac{\alpha}{\beta}\left(1 + \frac{x}{\beta}\right)^{-\alpha-1},$$

with form parameter $\alpha = 1.5$ and scale parameter $\beta = 0.5$.

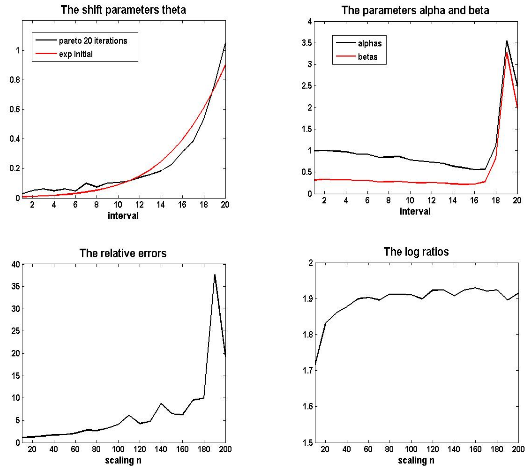
No exponential shifting possible.

Importance sampling with new densities Pareto(α_k, β_k) on subinterval I_k.

Cross-entropy algorithms: (i) updating both parameters; (ii) updating form parameters only; (iii) updating scale parameters only.

Results for (i).

Plots



Parameters after 20 iterations for scaling $n = 40$.

IS performance (20,000 runs) for scalings $10, \ldots, 200$ after CE iterations.

Outline

1. Model and problem
2. Large deviations
3. Importance sampling
4. Heuristics
5. Numerical results
6. Conclusion

Conclusion

- A rare event problem in the $M/G/\infty$ queue.
- Large deviations asymptotics.
- Importance sampling algorithm with cross-entropy improvement.
- Algorithm is ‘close’ to asymptotic optimal.