Importance sampling algorithms for first passage time probabilities in the infinite server queue

Ad Ridder

Department of Econometrics
Vrije Universiteit Amsterdam
aridder@feweb.vu.nl
http://staff.feweb.vu.nl/aridder/

RESIM Rennes
September 24-26, 2008

The $G/G/\infty$ queue

- i.i.d. interarrival times $U_1, U_2, \ldots$ with density function $f(x)$.
- i.i.d. service times $V_1, V_2, \ldots$ with density function $g(x)$, cdf $G(x)$, complementary cdf $\overline{G}(x) = 1 - G(x)$.
- Infinitely many servers: upon arrival service starts immediately.
- Finite means with rates λ and μ, resp.
- Light-tailed or heavy-tailed.

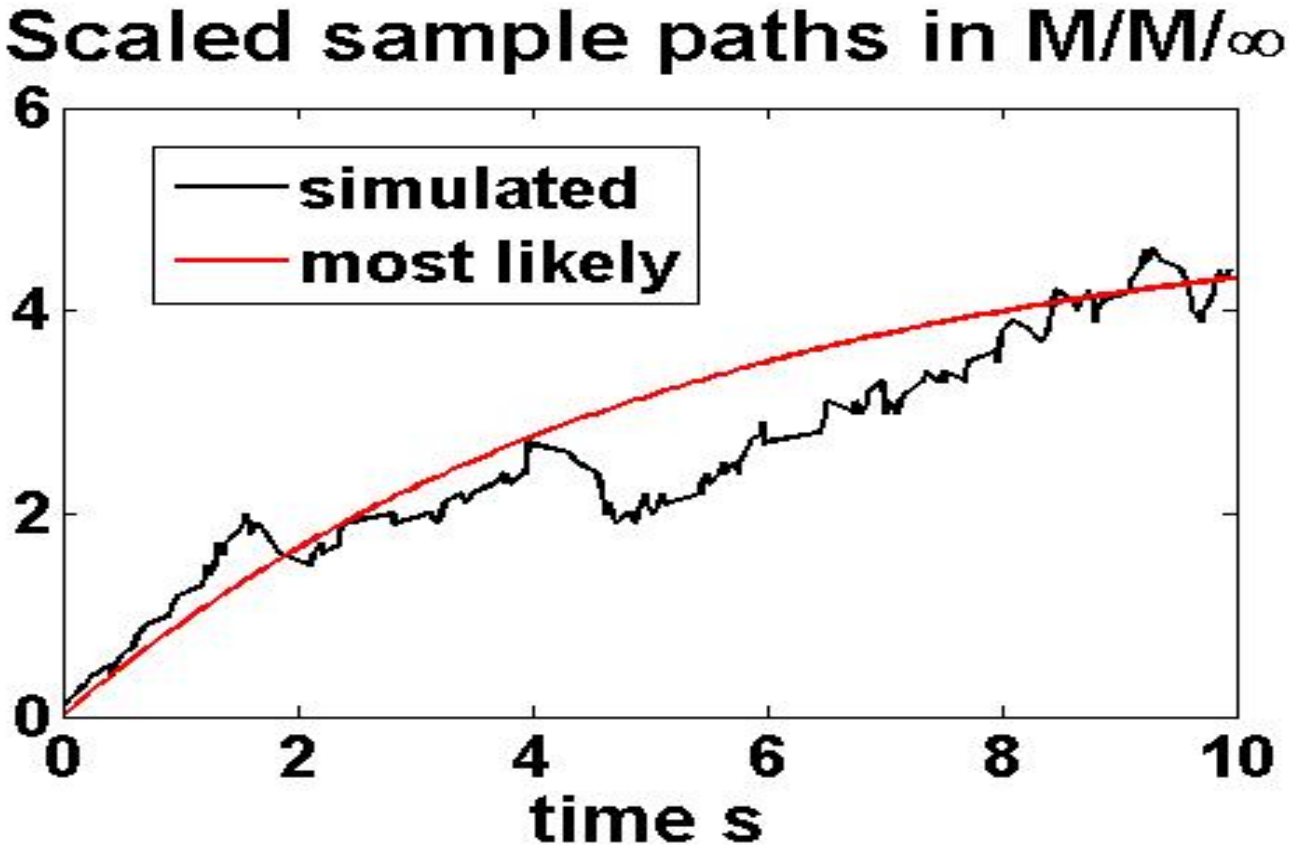
The rare event

- Scale interarrivals by n: $U_1/n, U_2/n, \ldots$.
- Let $Q_n(s)$ be the number of occupied servers at time s in the n-th system ($s \geq 0$, $n = 1, 2, \ldots$).
- Always $Q_n(0) = 0$.
- First passage times $T_n(j) = \inf\{s \geq 0 : Q_n(s) = j\}$.
- Target probability

$$\ell_n = P(T_n(nx) \leq t)$$

for some specified (fixed) time horizon $t > 0$ and large overflow level nx.

Typical sample paths don't reach the rare event



Data: $\lambda = 1, \mu = 0.2, x = 6, t = 10, n = 10$.

Large deviations

Theorem

$\lim_{n\to\infty} \frac{1}{n} \log \ell_n = -J(t,x).$

Here, $J(t,x)$ is the Legendre-Fenchel transform of the limiting cumulant generating function at time t:

$$J(t,x) = \sup_{\theta\in\mathbb{R}} \left(\theta x - \psi_Q(\theta,t)\right),$$

$$\psi_Q(\theta,t) = \lim_{n\to\infty} \frac{1}{n} \log E\left[e^{\theta Q_n(t)}\right].$$

Proof.

From Glynn (1995): large deviations for tail probabilities, i.e., for any $s > 0$

$$\lim_{n\to\infty} \frac{1}{n} \log P(Q_n(s)/n \geq x) = -J(s,x).$$

Show that $J(s,x)$ is decreasing in s, and apply the principle of the largest term:

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{n} \log \ell_n &= \lim_{n\to\infty} \frac{1}{n} \log\, P(T_n(nx) \leq t) \\
&= \lim_{n\to\infty} \frac{1}{n} \log P\left(\bigcup_{s\leq t} \{Q_n(s) \geq nx\} \right) \\
&= -\inf_{s\leq t} J(s,x) = -J(t,x).
\end{aligned}$$

□

Logarithmically efficient estimator

Target $\ell_n = P(A_n)$, where $\ell_n \to 0$ exponentially fast as $n \to \infty$.

Suppose Y_n is an unbiased estimator, $E[Y_n] = \ell_n$.

Definition

Estimator is logarithmically efficient if

$$\lim_{n\to\infty} \frac{\log E[Y_n^2]}{\log E[Y_n]} = 2. \tag{1}$$

Importance sampling algorithms

We have considered several ideas.

1. The optimal path approach:
   First apply large deviations to the $M/M/\infty$-model (Shwartz and Weiss 1995); identify the optimal path; change of measure so that is becomes the most likely path; adapt to deal with the general $G/G/\infty$.
2. Consider the algorithm for tail probabilities $P(Q_n(t)/n \geq x)$ for fixed t given in Szechtman and Glynn (2002); adapt to deal with first passage times.
3. An algorithm based on the cross-entropy method.

Comparison of the associated estimators

We consider three performance measures for estimators.

- RHW: relative half width of 95% confidence interval.
- RAT: estimated ratio (1).
- EFF: $-\log$(Variance estimator $\times$ used computer time).

Better performance when RHW is smaller, RAT is higher and EFF is larger.

1. The optimal path approach

Consider the $M/M/\infty$ model.

Under the change of measure the interarrival times and the service times have exponentially tilted densities with time-dependent tilting parameters that are all updated after each arrival or departure:

$$\begin{aligned} f^{\alpha}(u) &= \exp\left(\alpha u - \psi_U(\alpha)\right) f(u) \\ g^{\beta}(v) &= \exp\left(\beta v - \psi_V(\beta)\right) g(v). \end{aligned} \tag{2}$$

The tilting parameters α, β at time s are determined by

$$\psi_U'(\alpha) = e^{-\theta(s)}/\lambda; \quad \psi_V'(\beta) = e^{\theta(s)}/\mu,$$

with $\theta : [0, t] \to \mathbb{R}_{\geq 0}$ the tilting function obtained from the optimal path. In the simulation $\theta(s)$ is applied at jump times s.

Discussion

We can show that it results in a logarithmically efficient importance sampling estimator for the $M/M/\infty$ problem.

Easy to implement for memoryless distributions.

Otherwise (the general $G/G/\infty$ model): apply (2) only at arrival epochs for the arriving customer.

This is algorithm 1.

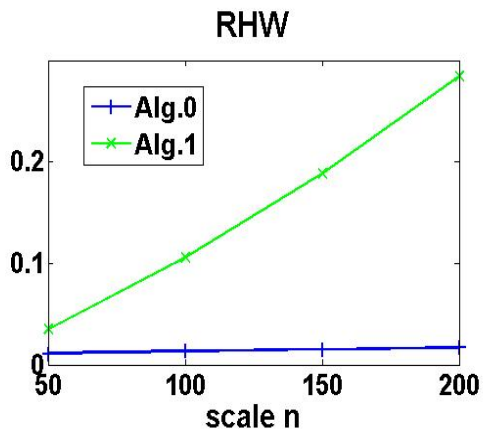
From experiments we found poor results when the service-time distribution is more variable than the exponential.

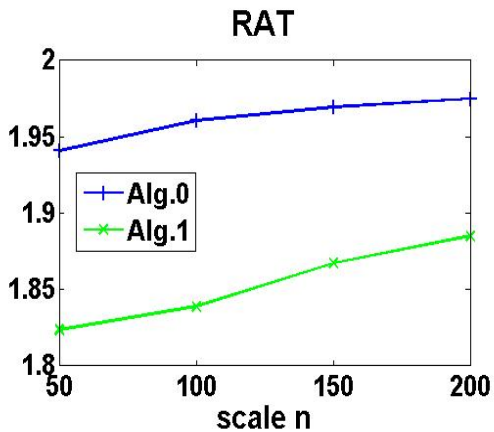
Not possible for heavy-tailed distributions.

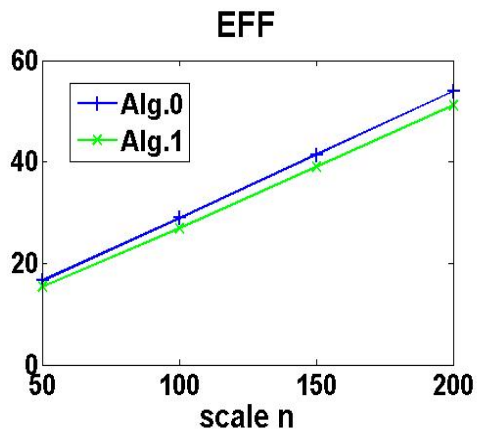
All and only

Comparison of the performance of the original IS estimator for $M/M/\infty$ with updating of all ongoing times after each event ('algorithm 0'), and its adapted counterpart associated with algorithm 1.

Averages of 5 repetitions with different seeds.







2. Adaptation of Szechtman-Glynn algorithm

The Szechtman-Glynn algorithm is a logarithmically efficient importance sampling algorithm for $P(Q_n(t)/n \geq x)$ with t fixed.

It is based on the same ideas as for the classical tail-problem $P(S_n/n \geq a)$ of partial sums of i.i.d. increments.

It simulates a change of measure that approximates

$$P^{\theta^*}(d\omega) = P(d\omega) \exp\left(n\left(\theta^* Q_n(t)/n - \psi_Q(\theta^*)\right)\right),$$

where $\theta^* > 0$ solves $\psi_Q'(\theta) = x$.

Work out $\psi'_Q(\theta) = x$ to obtain

$$\int_0^t \alpha(s)p(s)\,ds = x.$$

$\alpha(s)$ is the arrival rate at time s and $p(s)$ is the probability the an arrival at time s is still present at time t.

$\alpha(s)$ and $p(s)$ can be calculated numerically and implemented in an importance sampling algorithm.

Adapted Szechtman-Glynn algorithm for first passage times

Why adapt?
Answer: it may happen that $Q_n(t) < nx$, but possibly the process has reached nx before t.

Interarrival times are exponentially tilted versions (see (2)) with tilting parameter $\alpha(s)$.

Arriving customer receives service from a distribution $G^*(v)$ such that $P(V^* > t - s) = \overline{G^*}(t - s) = p(s)$.

This is algorithm 2.

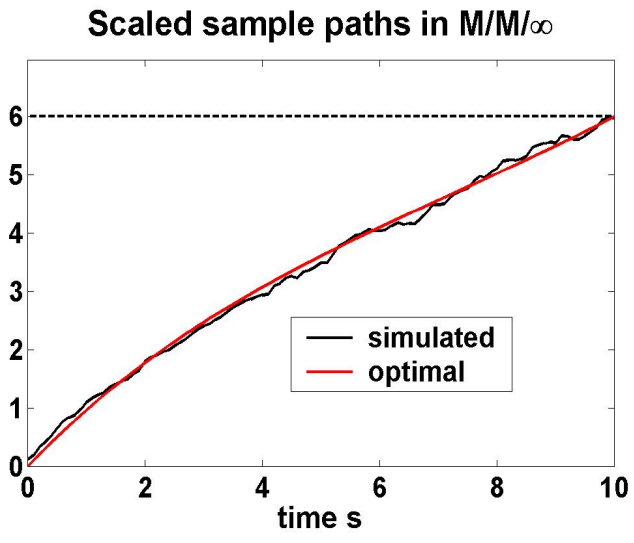
Discussion

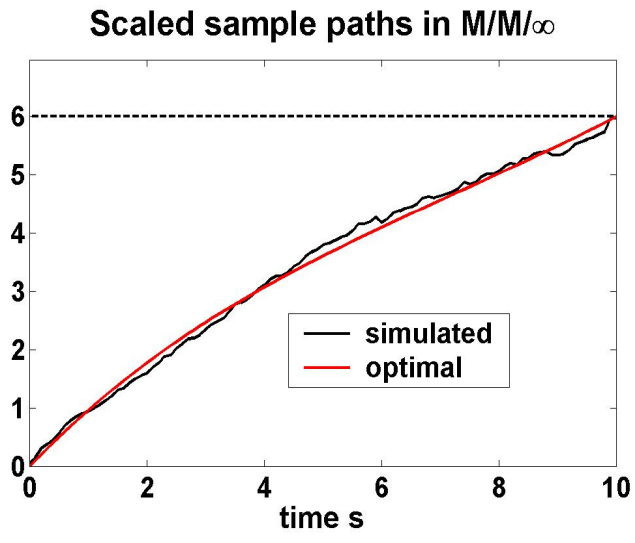
From experiments we found that the performance of (the estimator associated with) algorithm 2 is better than the performance of algorithm 1 when the distributions become more variable than the exponential, and otherwise it is worse.

Applicable for heavy-tailed distributions (both arrivals and services).

Illustration of simulated sample paths

Importance sampling algorithms 1 and 2 for $M/M/\infty$.





Data: $\lambda = 1, \mu = 0.2, x = 6, t = 10, n = 10$.
Average of 10 sample paths.
Left: algorithm 1. Right: algorithm 2.

3. Cross-entropy

A. For light tails.

Partition $[0, t]$ into M subintervals of equal size.

In the m-th subinterval, apply (2) at arrival epochs for the arriving customer, with tilting parameters α_m, β_m, i.e.,

$$f^{\alpha_m}(u) = \exp\left(\alpha_m u - \psi_U(\alpha_m)\right) f(u)$$
$$g^{\beta_m}(v) = \exp\left(\beta_m v - \psi_V(\beta_m)\right) g(v).$$

Use the cross-entropy method for finding 'optimal' tilting parameters.

B. For heavy tails.

Same approach, however, in the m-th subinterval take the importance sampling densities from the same parameterised families as the original densities.

Example: suppose the service time V is Pareto with form parameter $\kappa > 0$ and scale parameter $\gamma > 0$.

Then we take as importance sampling density on the m-th subinterval a Pareto density with form parameter κ_m and scale parameter γ_m.

Let the cross-entropy method find ‘optimal’ parameters.

Experiments

Data: $\lambda = 1, \mu = 0.2, x = 6, t = 10$.

1. $M/M/\infty$.
   Scaling $n = 50{:}50{:}200$, with $\ell_{200} \approx$ 5.4e-027.
   Sample size $k = 50000$.
   In cross entropy: at most 5 iterations of 7000 samples.
2. $M/\mathrm{Cox}_2/\infty$, with $c^2[V] = 4$.
   Scaling $n = 10{:}10{:}50$, with $\ell_{50} \approx$ 1.7e-028.
   Sample size $k = 50000$.
   In cross entropy: at most 20 iterations of 7000 samples.

NB, Coxian distribution with two phases:

$$V = \mathrm{Exp}(\mu_1) + B \cdot \mathrm{Exp}(\mu_2),$$

with B a Bernoulli(b) rv. Parameters μ_1, μ_2, b via two moment fit with gamma normalisation.

3. $\mathrm{Hyp}_2/M/\infty$, with $c^2[U] = 5$.
   Scaling $n = 100{:}100{:}500$, with $\ell_{500} \approx 1.0\text{e-}017$.
   Sample size $k = 50000$.
   In cross entropy: at most 15 iterations of 7000 samples.
4. $\mathrm{Hyp}_2/\mathrm{Par}/\infty$, with $c^2[U] = 5$ and $\mathrm{Var}[V] = \infty$.
   Scaling $n = 20{:}10{:}60$, with $\ell_{60} \approx 7.9\text{e-}016$.
   Sample size $k = 50000$.
   In cross entropy: at most 10 iterations of 5000 samples.

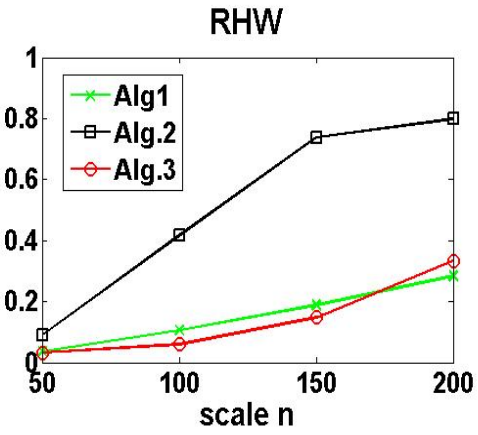
NB, Hyperexponential distribution with two phases:

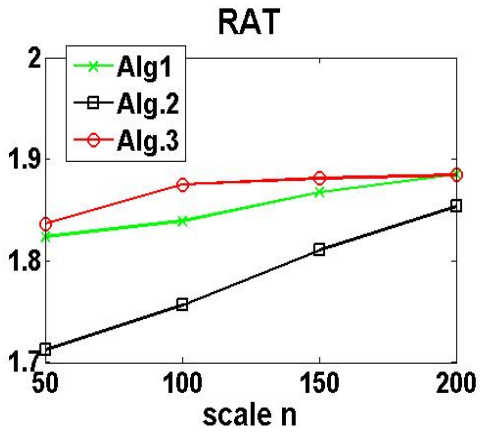
$$U = B \cdot \mathrm{Exp}(\mu_1) + (1 - B) \cdot \mathrm{Exp}(\mu_2),$$

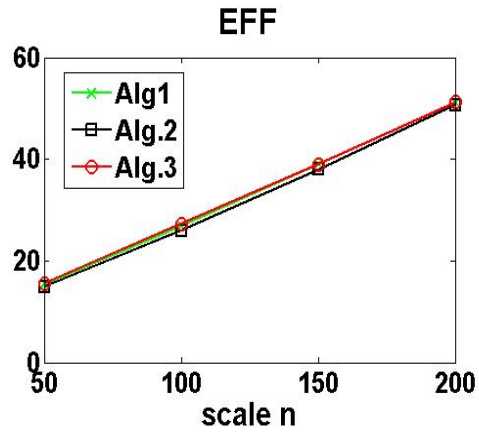
with B a Bernoulli(b) rv. Parameters μ_1, μ_2, b via two moment fit with balanced means.

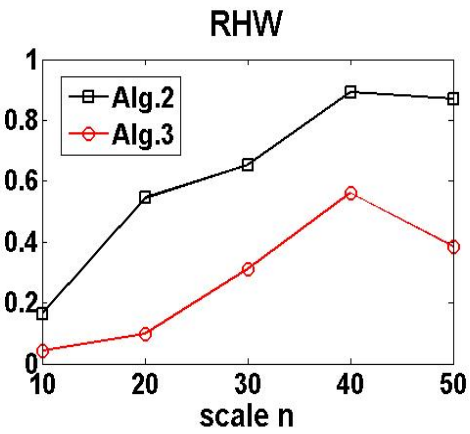
Results $M/M/\infty$ (top) $M/\mathrm{Cox}_2/\infty$ (bottom)

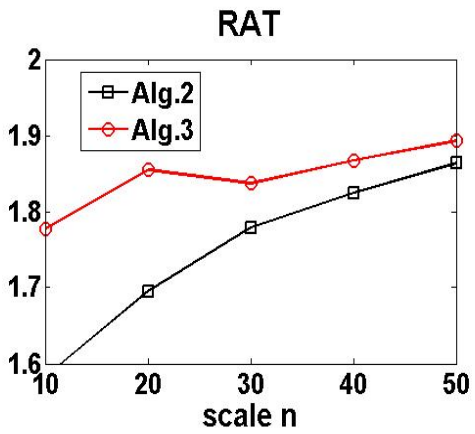
Averages of 5 repetitions with different seeds.

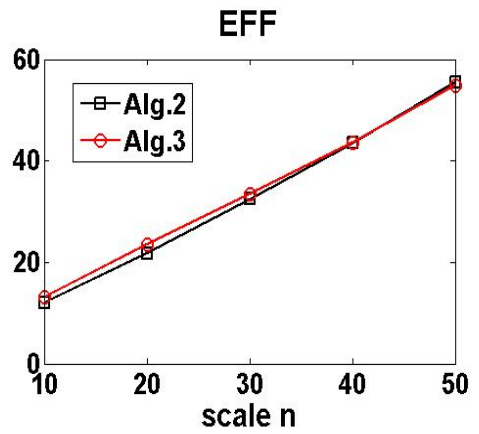




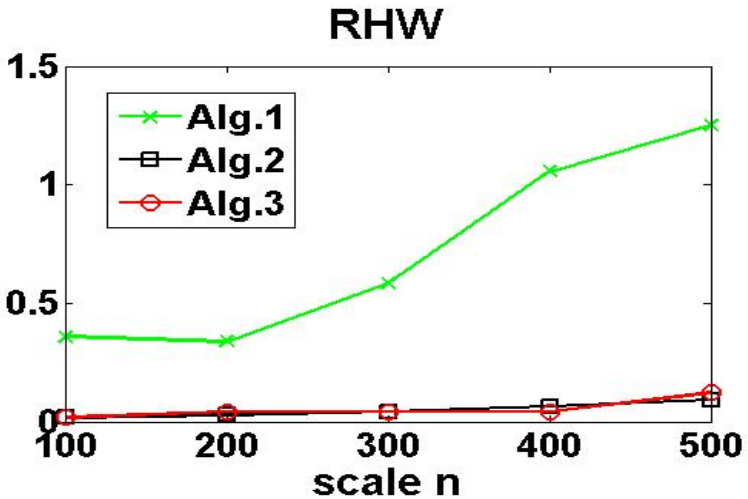


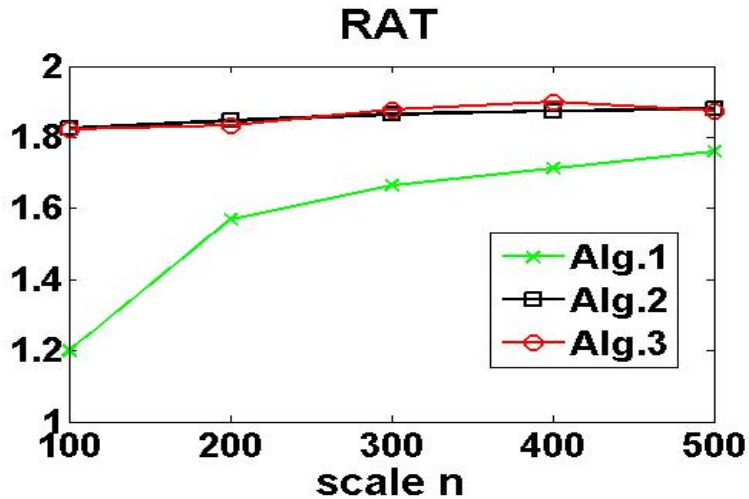


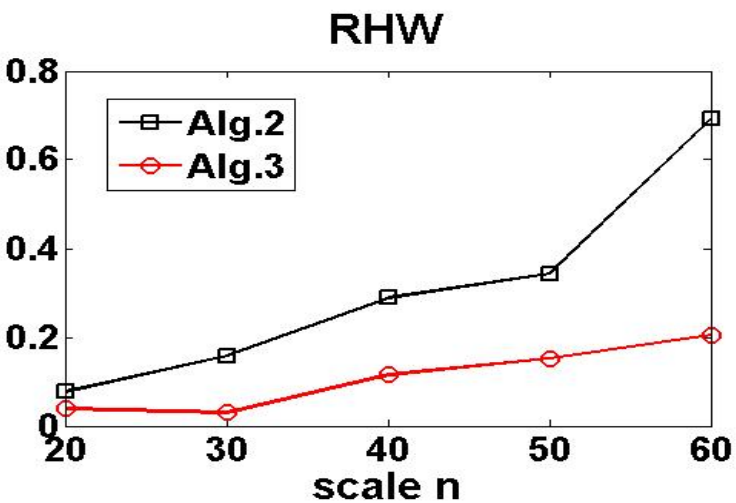


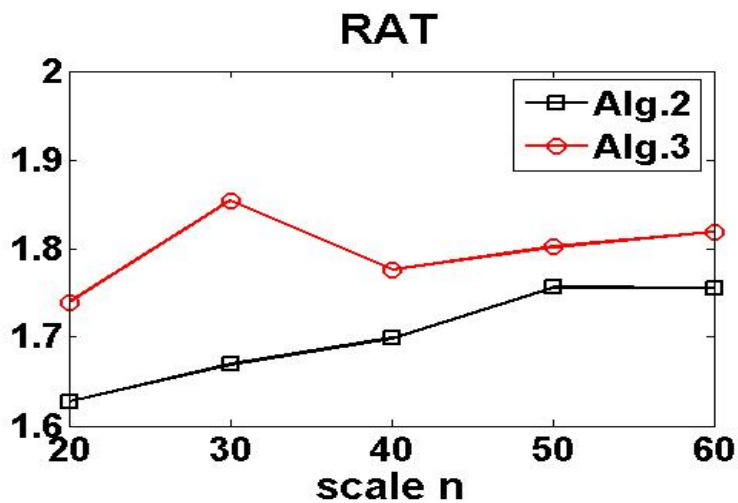


Results $\mathrm{Hyp}_2/M/\infty$ (top) $\mathrm{Hyp}_2/\mathrm{Par}/\infty$ (bottom)









Conclusion

- Rare event simulation in $G/G/\infty$ model.
- First passage time probabilities.
- Three importance sampling algorithms.
- Cross-entropy best algorithm in general but in specific cases might be second best.
- Further research: prove logarithmic efficiency of algorithms 2 and/or 3 in the general case.

References

1. Shwartz A., Weiss A., 1995.
   Large Deviations for Performance Analysis: Queues, Communications and Computing. Chapman Hall, London.
2. Glynn P., 1995.
   Large deviations for the infinite server queue in heavy traffic. In: Kelly F., Williams R. (Eds), Stochastic Networks, IMA Vol. 71. Springer, pp. 387-394.
3. Szechtman R., Glynn P., 2002.
   Rare-event simulation for infinite server queues. Proceedings of the 2002 Winter Simulation Conference, Vol. 1. IEEE Press, pp. 416-423.
4. Rubinstein R.Y., Kroese D.P., 2004.
   The cross-entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation and machine learning. Springer, New York.