Importance sampling simulation of the fork-join queue

Ad Ridder

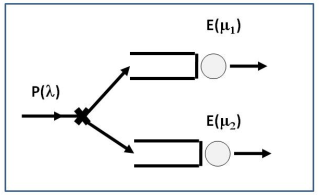
Department of Econometrics
Vrije Universiteit Amsterdam
`aridder@feweb.vu.nl`
`http://staff.feweb.vu.nl/aridder/`

PWS
July 1, 2009

The fork-join queue

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 $\lambda < \min(\mu_1, \mu_2)$.



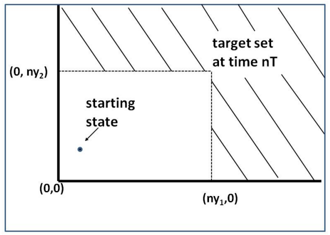
Folklore application: two bathrooms, one for men and one for women, and arrivals of couples.

Original motivation: machine with parallel processors (Hahn&Flatto 1984).

More general: allow individual arrivals (Wright 1992; Shwartz-Weiss book).

The rare event

$\{S(k) = (S_1(k), S_2(k)) : k = 0, 1, \ldots\}$
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:

$$\gamma_n(x, y, T) = \mathbb{P}(S_1(nT) \geq ny_1 \text{ or } S_2(nT) \geq ny_2 | S(0) = nx),$$

for fixed scaled initial state $x = (x_1, x_2) \in \mathbb{R}^2_+$, fixed scaled threshold $y = (y_1, y_2) \in \mathbb{R}^2_+$, fixed scaled horizon $T > 0$, and parameter $n \to \infty$.

The rarity set

The scaled set of interest:

$$D = \{\eta \in \mathbb{R}^2_+ : \eta_1 \geq y_1 \text{ or } \eta_2 \geq y_2\},$$

i.e.

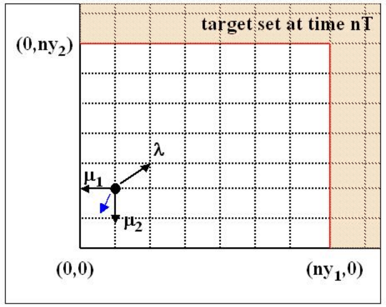
$$\gamma_n(x, y, T) = \mathbb{P}\left(S(nT)/n \in D | S(0) = nx\right).$$

The difficulty for importance sampling is twofold:

(i). the rarity set is not convex (Dupuis&Wang 2007);

(ii). the rarity set cannot decomposed in two disjoint sets such that the separate probabilities are estimated by efficient importance sampling estimators (Glassermann&Wang 1997).

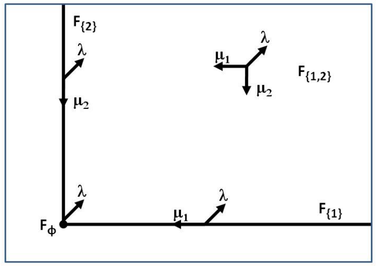
A sample path of the fork-join queue



The blue arrow indicates the ‘natural’ drift.
We show the transition rates; for the discrete-time Markov chain these are normalized to probabilities.

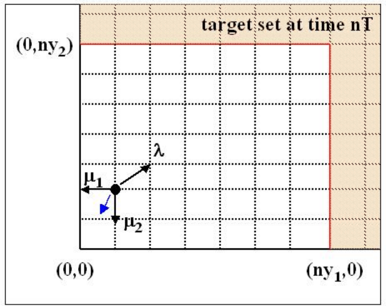
Face-homogeneous random walk

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_Λ.
We might associate a random walk $(S_\Lambda(k))_{k=0}^{\infty}$ with jump variable X_Λ with probabilities $p_\Lambda(j) \doteq p_{s,s+j}$.

A sample path after change of measure



The idea of the importance sampling scheme: until a certain time $n\tau$ it follows the original transition probabilities.

Importance sampling scheme

Our importance sampling scheme will be a mixture of two sets of exponentially shifted jump probabilities of the jump variables X_Λ.

For any $\theta \in \mathbb{R}^2$, the θ-shifted jump X_Λ^θ has jump probabilities

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

where $\psi_\Lambda(\cdot)$ is the log moment generating function of jump variable X_Λ.

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.

Sample path large deviations

Define continuous processes $(S^{[n]}(t))_{0\le t\le T}$, $n = 1, 2, \ldots$, by scaling:

$$S^{[n]}(t) = S(nt)/n \text{ for } t = 0, 1/n, 2/n \ldots, T,$$

and linear interpolation in the other points.

Consider absolute continuous functions $\phi : [0, T] \to \mathbb{R}^2_+$. Then (Ignatiouk 2005)

$$-\lim_{\epsilon\downarrow 0}\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}\left(\sup_{0\le t\le T}\left|S^{[n]}(t)-\phi(t)\right|<\epsilon\right)=\int_0^T \ell_{\Lambda(\phi(t))}(\phi'(t))\,dt \doteq I(\phi),$$

where $\ell_\Lambda : \mathbb{R}^2 \to [0, \infty]$ are so-called locate rate functions (see forthcoming slides).

Sample path large deviations (cont'd)

Notice that for the fork-join queue problem

$$\gamma_n(x, y, T) = \mathbb{P}\left(S(nT)/n \in D|S(0)/n = x\right) = \mathbb{P}\left(S^{[n]} \in E\right),$$

where E is an appropriate set of absolute continuous paths $\phi : [0, T] \to \mathbb{R}^2_+$ with specifically $\phi(0) = x$ and $\phi(T) \in D$ (the rarity set).

Let $\tilde{E} \subset E$ be the subset of piecewise linear paths of the following form.

> $\phi = \phi_{\tau,v}$: *it follows the natural drift until time τ and then it goes straight at constant speed $v = \phi'(t)$ to a point in the rarity set D.*

One can show that

$$\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}\left(S^{[n]}\in E\right)=\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}\left(S^{[n]}\in\tilde{E}\right)=-I(\tilde{E}),$$

where $I(\tilde{E}) = \inf_{\tau,v} I(\phi_{\tau,v})$, and $I(\phi_{\tau,v}) = (T-\tau)\ell_\Lambda(v)$ assuming that the second part of the path runs entirely in face F_Λ.

The local rate functions

The local rate function $\ell_\Lambda(v)$ is the convex conjugate of a certain convex function $\psi : \mathbb{R}^2 \to \mathbb{R}$:

$$\ell_\Lambda(v) = \sup_{\theta\in\mathbb{R}^2}\left(\langle\theta, v\rangle - \psi(\theta)\right).$$

It can be determined numerically via the method in (Ignatiouk 2001).

The optimizer θ_v is called the optimal shift factor associated with speed v.

When θ_v is used to exponentially shift the jump probabilities of the internal jump variable $X_{\{1,2\}}$, the speed vector v corresponds with the drift of the shifted jump variable $X^{\theta_v}_{\{1,2\}}$, thus restricted for being a convex combination of the jumps $(-1, 0), (0, -1), (1, 1)$.

Clearly, the speed vectors in the boundary faces $F_{\{1\}}$ and $F_{\{2\}}$ are restricted to be 0 in the perpendicular direction and between -1 and 1 in the parallel direction. (They do not correspond with drifts!)

Denote by V_Λ the set of feasible speed vectors v in face F_Λ.

The paths with constant speed

We restrict to starting state $x = (0, 0)$.

Consider paths that stay in 0 during τ time units.

Let $V_\Lambda(\tau)$ be the set of feasible speed vectors v in face F_Λ such that $(T-\tau)v \in D$, i.e.,

$$V_\Lambda(\tau) = \{v \in V_\Lambda : \phi_{\tau,v} \in \tilde{E}\}.$$

Hence, there is a one-to-one correspondence

$$\tilde{E} \leftrightarrow \bigcup_{\tau\ge 0} V_{\{1,2\}}(\tau) \cup \bigcup_{\tau\ge 0} V_{\{1\}}(\tau) \cup \bigcup_{\tau\ge 0} V_{\{2\}}(\tau)$$
$$= V_{\{1,2\}}(0) \cup V_{\{1\}}(0) \cup V_{\{2\}}(0) \doteq V(0).$$

Efficient importance sampling

The idea is to choose (better: find) a set of speeds $v^{(1)}, \ldots, v^{(m)}$ and associated optimal shift factors $\theta^{(1)}, \ldots, \theta^{(m)}$, such that

$$V(0) \subset \bigcup_{i=1}^{m} \mathcal{H}(v^{(i)}),$$

where

$$\mathcal{H}(v) = \{w \in V(0) : \langle \theta_v, w \rangle \geq \langle \theta_v, v \rangle\}.$$

Then, any mixture importance sampling scheme with exponentially shifted probability densities using shift factors $\theta^{(1)}, \ldots, \theta^{(m)}$ is asymptotically optimal (Bucklew 1990, 2004).

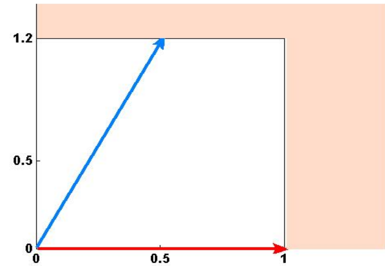
Example

$\lambda = 1, \mu_1 = 1.5, \mu_2 = 2, x = (0,0), y = (1, 1.2), T = 10.$

We were able to find a solution of two speed vectors such that

$$V(2) \doteq \bigcup_{\tau \geq 2} V_{\{1,2\}}(\tau) \cup \bigcup_{\tau \geq 2} V_{\{1\}}(\tau) \cup \bigcup_{\tau \geq 2} V_{\{2\}}(\tau) \subset \mathcal{H}(v^{(1)}) \cup \mathcal{H}(v^{(2)}).$$

We mix them 0.8 (red path) and 0.2 (blue path).

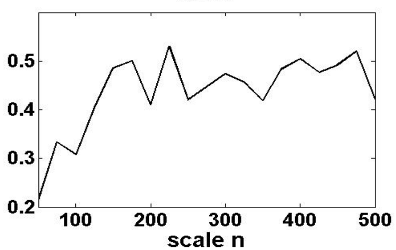


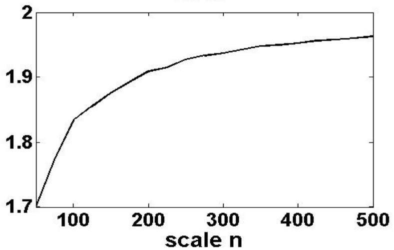
Example (cont'd)

Results for scalings $n = 25$–500 with sample size $k = 50000$ for plotting the relative half width of the 95% confidence interval for estimator $\widehat{\gamma}_n$,

RHW = $1.96\sqrt{\mathrm{Var}[\widehat{\gamma}_n]}/\mathbb{E}[\widehat{\gamma}_n]$,

and ratio RAT $= \log \mathbb{E}[(\widehat{\gamma}_n)^2]/\log \mathbb{E}[\widehat{\gamma}_n]$.
(efficient estimators have RAT that converge to 2).





Conclusion and further research

- We have developed an importance sampling scheme which is a mixture of two time-dependent (state-independent) exponentially shifted densities.
- Excellent simulation results.
- Need to prove that using $V(2)$ in stead of $V(0)$ still gives asymptotical optimality. (The large deviations asymptotic still holds.)
- Further investigations include other starting points, and other algorithms, for instance with mixing transition probabilities that depend on state and time.

References

1. Flatto, L., and Hahn, S., 1984. Two parallel queues created by arrivals with two demands, *SIAM Journal on Applied Mathematics* 44, pp. 1041-1053.
2. Wright, P.E., 1992. Two parallel processors with coupled inputs, *Advanced of Applied Probability* 24, pp. 986-1007.
3. Shwartz, A., and Weiss, A., 1995. *Large Deviations for Performance Analysis*, Chapman & Hall.
4. Dupuis, P., and Wang, H., 2007. Subsolutions of an Isaacs equation and efficient schemes for importance sampling. *Mathematics of Operations Research* 32, pp. 723-757.
5. Glassermann, P., and Wang, Y., 1997. Counterexamples in importance sampling for large deviations probabilities, *Annals of Applied Probability* 7, pp. 731-746.
6. Ignatiouk-Robert, I., 2001. Sample path large deviations and convergence parameters, *Annals of Applied Probability* 11, pp. 1292-1329.
7. Ignatiouk-Robert, I., 2005. Large deviations for processes with discontinuous statistics, *Annals of Probability* 33, pp. 1479-1508.
8. Sadowsky, J.S., and Bucklew, J.A., 1990. On large deviations theory and asymptotically efficient Monte Carlo estimation. *IEEE Transactions on Information Theory* 36, pp. 579-588.