Monte-Carlo Methods for the Estimation of Rare Event Probabilities

Kevin Leder

December 2, 2011

Outline

Estimation of Small Probabilities via Monte Carlo

- Why try to estimate the probability of rare events? Aren't they just 0?
- Phrase in terms of probabilities and random variables, there is a random variable Z and a set A such that $\mathbb{P}(Z \in A) \approx 0$.
- Useful to embed rare event into sequence of rare events and study asymptotic properties i.e., consider the events $\{X^n \in A\}$.
- We are interested in setting where the probabilities decay exponentially, i.e. there exists a $\gamma > 0$ such that

$$\lim_{n \to \infty} \frac{1}{n} \log \mathbb{P}\left(Z^n \in A\right) = -\gamma < 0.$$

Estimating Rare Event Probabilities via Standard Monte Carlo

- Suppose we are interested in estimating $\mathbb{P}(Z^n \in A)$ for some fixed n.
- For a large integer k draw an i.i.d. vector $(Z_1^n, \ldots, Z_k^n)$ then form the estimator
$$p_{n,k} = \frac{1}{K} \sum_{j=1}^{k} 1_A(Z_j^n),$$
which is unbiased and consistent.
- Consider relative error of estimator though
$$\text{RE}(p_{n,k}) = \frac{\text{sd}(p_{n,k})}{\mathbb{E}[p_{n,k}]} = \sqrt{\frac{1 - \mathbb{P}(Z^n \in A)}{k\mathbb{P}(Z^n \in A)}} \approx \frac{1}{\sqrt{k\mathbb{P}(Z^n \in A)}}.$$
- Number of replications k has to grow with $1/\mathbb{P}(X^n \in A)$ to keep relative error bounded.

Two Solutions

- Importance sampling: Simulate system under alternative dynamics so that event of interest is no longer rare. Keep track of likelihood ratio so that you can renormalize final answer to create unbiased estimator.
- Particle based methods: Simulate lots of correlated copies of system under original dynamics, these methods can be viewed as a type of branching random walk.

Importance Sampling

Importance Sampling

Importance Sampling

- For estimating $p_n = \mathbb{P}(Z^n \in A)$, first construct new sampling measure $\mathbb{Q}$ then form estimator by averaging independent replications of
$$\hat{p}_n = \frac{d\mathbb{P}}{d\mathbb{Q}} 1_A(Z^n),$$
where Z^n is sampled according to measure $\mathbb{Q}$.
- Judge the performance of $\hat{p}_n$ via its variance, (or equivalently 2nd moment)
$$\mathbb{E}^{\mathbb{Q}}[\hat{p}_n^2] = \mathbb{E}[\hat{p}_n].$$
- In order to control relative error would like *strong efficiency*
$$\sup_{n<\infty} \frac{\mathbb{E}[\hat{p}_n]}{p_n^2} < \infty.$$
- This criteria is difficult to achieve in practice so settle for *logarithmic efficiency*
$$\liminf_n -\frac{1}{n} \log \mathbb{E}[\hat{p}_n] \geq 2\gamma.$$

Importance Sampling for Random Walks

- Consider estimating $p_n(A) = P(Z_n/n \in A)$ where $Z_n = X_1 + \ldots + X_n$ and $\{X_i\}$ is an i.i.d sequence of d dimensional random vectors that satisfy

$$\psi(\theta) = \log \mathbb{E}[e^{\langle \theta, Y \rangle}] < \infty,$$

  for θ in a neighborhood of the origin. Assume that $E[X_1] \notin \bar{A}$.
- Create a sampling measure by exponential tilt of each increment of X_i, for each $\alpha \in \mathbb{R}^d$ define a sampling measure

$$\mathbb{Q}_\alpha(X_1 \in dx_1, \ldots, X_n \in dx_n) = \frac{e^{\langle \alpha, x_1 \rangle}}{e^{\psi(\alpha)}} \mathbb{P}(X_1 \in dx_1) \cdots \frac{e^{\langle \alpha, x_n \rangle}}{e^{\psi(\alpha)}} \mathbb{P}(X_n \in dx_n).$$

- For this change of measure denote our estimator of $p_n(A)$ by $\hat{p}_n(A, \alpha)$, then 2nd moment single replication of estimator is

$$\mathbb{E}[\hat{p}_n(A, \alpha)] = \mathbb{E}[1_{\{Z_n/n \in A\}} e^{n(\psi(\alpha) - \langle \alpha, Z_n/n \rangle)}].$$

- Study asymptotic properties of 2nd moment via large deviation theory.

Large Deviations Principle

- A sequence of random variables $\{Z^n\}$ taking values in a Polish space X satisfy a large deviations principle (LDP) with rate function $I : X \to [0, \infty]$ if
  1. I has compact level sets
  2. For every Borel set $A \subset X$

$$
\begin{aligned}
- \inf_{x \in A^\circ} I(x) &\le \liminf_{n \to \infty} \frac{1}{n} \log \mathbb{P}(Z^n \in A) \\
&\le \limsup_{n \to \infty} \frac{1}{n} \log \mathbb{P}(Z^n \in A) \le - \inf_{x \in \bar{A}} I(x)
\end{aligned}
$$

- A useful alternative formulation: for any bounded and continuous $f : X \to \mathbb{R}$ the following holds

$$
\lim_{n \to \infty} -\frac{1}{n} \log \mathbb{E} e^{-nf(Z^n)} = \inf_{x \in X} [f(x) + I(x)].
$$

Large Deviations for Random Walks

- Suppose that $Z_n = X_1 + \ldots + X_n$ for an i.i.d sequence $\{X_i\}$ of d dimensional vector that satisfy

$$\psi(\theta) = \log \mathbb{E}[e^{\langle \theta, Y \rangle}] < \infty,$$

for θ in a neighborhood of the origin.

- Then Z_n/n satisfies an LDP with rate function (Cramer's Theorem)

$$I(\beta) = \sup_{\alpha \in \mathbb{R}^d} [\langle \alpha, \beta \rangle - \psi(\alpha)].$$

- In the 1-d setting, if $a > E[X_1]$ then Cramer's theorem gives that

$$P(Z_n > na) = e^{-n(I(a)+o(1))},$$

where $I(a) = \inf_{x>a} I(x) = I(a) = a\theta_a - \psi(\theta_a)$ and θ_a solves $\psi'(\theta_a) = a$.

Using Large Deviations for Importance Sampling

- Using Cramer's Theorem and alternative formulation of LDP we can approximate 2nd moment of IS estimator for large n,

$$\frac{1}{n}\log \mathbb{E}[1_{\{Z_n/n\in A\}} e^{n(\psi(\alpha)-\langle\alpha,Z_n/n\rangle)}] \approx -\inf_{x\in A}[I(x)-\psi(\alpha)+\langle\alpha,x\rangle].$$

- The goal of importance sampling is to minimize variance which gives following minmax problem

$$\sup_{\alpha\in\mathbb{R}^d}\inf_{x\in A}[I(x)-\psi(\alpha)+\langle\alpha,x\rangle].$$

- If A is convex then

$$\sup_{\alpha\in\mathbb{R}^d}\inf_{x\in A}[I(x)-\psi(\alpha)+\langle\alpha,x\rangle] = \inf_{x\in A}\sup_{\alpha\in\mathbb{R}^d}[I(x)-\psi(\alpha)+\langle\alpha,x\rangle]$$

$$= 2\inf_{x\in A} I(x).$$

- Which $\alpha\in\mathbb{R}^d$ to use? Let $x^* = arginf_{x\in A} I(x)$ then we use change of measure defined by using tilte α_{x^*} which is solution of $D\psi(\alpha_{x^*}) = x^*$.
- If A is convex, this is a logarithmically efficient estimator.

Importance of Convexity

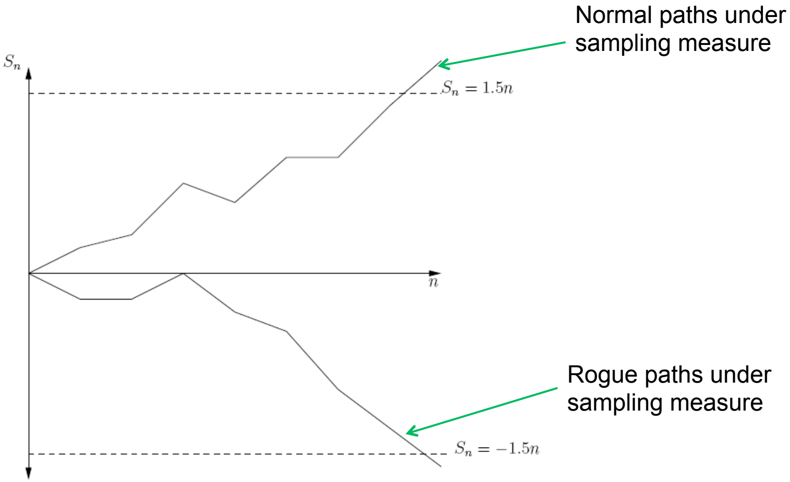
- Glasserman and Yang (97) consider the problem of estimating $P(|Z_n| > 1.5n)$ where the increments $X_i = A_i - B_i$ and $A_i \sim N(1.5, 1)$ and $B \sim \exp(1)$.
- If set were convex then we need to find $x^* = arginf_{x:|x|>1.5} I(x) = 1.5$, then use change of measure based on $\alpha_{1.5}$ i.e.,

$$\frac{d\mathbb{P}}{d\mathbb{Q}}(x_1) = e^{\alpha_{1.5}x_1 - \psi(\alpha_{1.5})}.$$

- However by pretending target set was convex we end up with terrible estimator

$$\limsup_{n\to\infty} \mathbb{E}[\hat{p}_n(A, \alpha_{1.5})] = \infty.$$

What went wrong?



A procedure for non-convex A

- Dupuis and Wang showed that for non-convex A logarithmic efficiency requires state-dependent changes of measure.
- Suppose that $A = A_1 \cup \ldots \cup A_m$, where A_j are closed convex sets. Define $y_j = arginf_{j \in A_j} I(y)$ and $\alpha_j \doteq \alpha_{y_j}$. Then a logarithmically efficient change of measure is given by using transition kernel

$$\mathbb{Q}(X_i \in dx | Z_{i-1} = z) = \sum_{j=1}^{m} r_i^j(z) e^{\langle \alpha_j, x_i \rangle - \psi(\alpha_j)} \mathbb{P}(X_i \in dx_i),$$

- The state-dependence mixture probabilities are described as follows

$$r_i^j(z) = \frac{w_i^j(z)}{\sum_{k=1}^m w_i^k(z)},$$

where

$$w_i^k(z) = \exp\left[n\langle \alpha_k, z - y_k \rangle + (n-i)\psi(\alpha_k)\right].$$

Importance Sampling in Finance

- Many option pricing problems can be viewed as rare-event calculations. Option only has value on a small set of sample space, so expected value is dominated by values on a rare set.
- Glasserman et al (99) looked at using importance sampling as a computational tool for pricing a variety of path dependent options. In setting of concave payoff function they present logarithmically efficient procedure for pricing options.
- Gausoni and Robertson extended framework of the Glasserman paper to a continuous time setting and establish that optimal change of measure in continuous time setting can be found by solving Euler-Lagrange equation. They assume that payoff function is concave.
- Several works by Glasserman have considered use of importance sampling for estimating value at risk, and conditional value at risk.
- Dupuis and Wang show that under very weak conditions on the payoff functional adaptive importance sampling can be used to evaluate option prices with logarithmic efficiency.

Splitting Method

Particle Based Methods

Splitting Method

- Will focus on a specific particle method called 'splitting method', first developed in Villen-Altamirano and Villen-Altamirano '94 who called it RESTART.
- Dean and Dupuis '08 presented a procedure for construction of efficient and stable splitting schemes. Will follow their notation.
- Model problem: X^n a sequence of stochastic processes on domain $D \subset \mathbb{R}^d$, and two disjoint sets A and B, define the sequence of stopping times $\tau_n = \min\{i : X^n(i) \in A \cup B\}$
- Goal estimate the probabilities

$$p_n(x) = \mathbb{P}\left(X^n(\tau_n) \in B | X^n(0) = x\right).$$

- Assume that there is a non-negative measurable function L such that

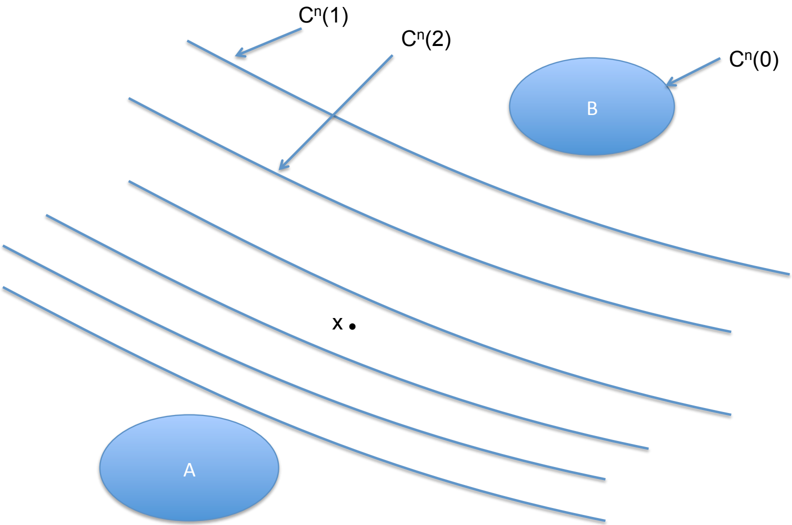
$$\lim_{n\to\infty} -\frac{1}{n} \log p_n(x)$$
$$= \inf\{\int_0^t L\left(\phi(s), \dot{\phi}(s)\right) ds : \phi(0) = 0, \phi(t) \in B, \phi(s) \in A^c \text{ for all } s \leq t\}.$$

The Splitting Algorithm

- Consider collection of nested sets $B = C^n(0) \subset C^n(1) \subset \ldots \subset C^n(M_n)$

1. Initiate simulation procedure with a single particle starting from position $x \in C^n(k)$ for some $k \geq 1$. Let $w_1 = 1$ initial weight associated to particle.
2. Evolve initial particle according to original transition kernel until either it hits A (dies) or level $C^n(k-1)$. If it hits $C^n(k-1)$ it is replaced by r identical particles ($r > 1$). Weight of descendant particles is weight of parent particle $\times 1/r$.
3. Procedure from step 1 is replicated for each descendant particle, carrying over the value of the weights at each level for the surviving particles.
4. Steps 1 to 3 are repeated until all particles have either died or reached level $C^n(0) = B$.

Splitting Method



The Splitting Estimator

- Consider collection of nested sets $B = C^n(0) \subset C^n(1) \subset \ldots \subset C^n(M_n)$ (Note will want $M_n = c * n$ for some $c > 0$).
- Nested sets are based on level sets of an 'importance function', U. Specifically define $L_z = \{y \in D : U(y) \leq z\}$ then

$$C^n(j) = L_{(j-1)/n}.$$

- An important function is the level function

$$\ell^n(y) = \min\{j \geq 0 : y \in C^n(j)\},$$

- Estimator for $p_n(x)$ is

$$R_n(x) = N_n(x)/r^{\ell_n(x)},$$

where $N_n(x)$ is number of particles that made it to B.

Analysis of Splitting Estimators

- For numerical stability want $E[N_n(x)] \approx r^{\ell_n(x)} p_n(x)$ to grow subexponentially, or $r^{\ell_n(x)} p_n(x) = \exp(o(n))$.
- For logarithmically optimal 2nd moment require that $r^{-\ell_n(x)} = p_n(x) \exp(o(n))$.
- Suppose we have a function $W(x)$ such that

$$p_n(x) \leq \exp\left(-nW(x) + o(n)\right),$$

  then it suffices to establish that

$$\ell_n(x) \log r - nW(x) = o(n)$$

- It's easy to see that $\ell_n(x) = \lceil nU(x) \rceil$ therefore we choose our importance function as $U(x) = W(x)/\log(r)$.
- See Dean and Dupuis for details.

Jackson Networks

Performance Comparison: Overflow in Jackson Networks

Open Jackson Networks

- Consider a network of d stations. Customers arrive to the network with arrival rate $\lambda = (\lambda_1, \ldots, \lambda_d)^T$, and the service rate of the d stations is encoded by $\mu = (\mu_1, \ldots, \mu_d)^T$.
- A job that leaves station i joins station j with probability $P_{i,j}$, and leaves the system with probability

$$P_{i,0} = 1 - \sum_{j=1}^{d} P_{i,j},$$

  this is called the routing matrix.
- We are interested in stable open Jackson networks, that is
  i) $\forall i$, either $\lambda_i > 0$ or $\lambda_{j_1} P_{j_1 j_2} \ldots P_{j_k i} > 0$ for some $j_1, \ldots, j_k$.
  ii) $\forall i$, either $P_{i0} > 0$ or $P_{ij_1} P_{j_1 j_2} \ldots P_{j_k 0} > 0$ for some $j_1, \ldots, j_k$.
  iii) The network is stable (i.e. a stationary distribution exists).

Basic Properties of Jackson Networks

- Assume without loss of generality that $\sum_{j=1}^{d}(\lambda_j + \mu_j) = 1$.
- Under the stability assumption the following

$$\phi_i = \lambda_i + \sum_{j=1}^{d} \phi_j P_{ji}, \qquad \forall i = 1, 2, ..., d$$

  has a unique solution $\phi^T = \lambda^T (I - P)^{-1}$.
- The traffic intensity at station i is in equilibrium is given by $\rho_i = \phi_i / \mu_i \in (0, 1)$.
- Define $\rho_* = \max_{1 \leq i \leq d} \rho_i$, and then set $\beta = |\{i : \rho_i = \rho_*\}$.
- Study system through embedded discrete time Markov chain $Q = \{Q(k) : k \geq 0\}$ where $Q(k) = (Q_1(k), \ldots, Q_d(k))$, and $Q_i(k)$ represents number of customers in station i immediately after kth transition.

Overflow Probabilities in Jackson Networks

- Consider a subset of stations encoded by the vector v, denote the total population in this subset by $N_v(x) = \langle x, v \rangle$.
- Will be interested in the following probability:

$$p_n^v = \mathbb{P}\{\text{ total population in stations encoded by } v \text{ reaches } n \text{ before returning to 0, starting from 0}\}.$$

- Can also define p_n^v via stopping times

$$T_{\{x\}} \triangleq \inf\{k \geq 1 : Q(k) = x\},$$
$$T_n^V \triangleq \inf\{k \geq 1 : N_v(Q(k)) \geq n\}.$$

- If we define $\mathbb{P}(\cdot) \doteq \mathbb{P}(\cdot|Q(0) = x)$ then

$$p_n^V = \mathbb{P}_0(T_n^V \leq T_{\{0\}}).$$

or more generally

$$p_n^V(x) = \mathbb{P}_x(T_n^V \leq T_{\{0\}}).$$

Dynamics of Q

- Queue length process is just a state-dependent random walk

$$Q(k+1) = Q(k) + \zeta\left(Q(k), Y(k+1)\right),$$

- ζ is a reflection function that prevents the queue-length process from taking negative values.
- The noise term $Y(k)$ represents outcome of next transition and has following pdf

$$\mathbb{P}\left(Y(k) = w\right) = \begin{cases} \lambda_i & \text{arrival at station } i, \\ \mu_i P_{ij} & \text{dep. at station } i \text{ goes to station } j, \\ \mu_i P_{i0} & \text{dep. at station } i \text{ leaves sys.} \end{cases}$$

Logarithmic Asymptotics of Overflow Probabilities in Jackson Networks I

- Large deviations theory dictates the existence of a function W

$$p_n^V(x/n) = \exp(-nW_V(x/n) + o(n)).$$

- By looking at Q/n we have the following via formal Taylor expansion

$$\begin{aligned}
1 &= \frac{1}{p_n^V(x/n)} \mathbb{E}\left[p_n^V(x/n + \frac{1}{n}\zeta(x/n, Y(1))) \right] \\
&\approx \mathbb{E} \exp\{-nW_V[x/n + \frac{1}{n}\zeta(x/n, Y(1))] + nW_V(x/n)\} \\
&= \mathbb{E} \exp\{-\partial W_V(x/n)^T \zeta(x/n, Y(1)) + o(1)\} \\
&= \exp(\psi(x/n, -\partial W_V(x/n)) + o(1)),
\end{aligned}$$

where $\psi(x, \theta) = \log \mathbb{E}\left[\exp\left(\theta^T \zeta(x, Y(k))\right)\right]$.

Logarithmic Asymptotics of Overflow Probabilities in Jackson Networks II

- In order to characterize logarithmic asymptotics of $p_n(x/n)$ need to find function W_V that satisfies

$$\psi(x/n, -\partial W_V(x/n)) = 0,$$

  or for an asymptotic logarithmic upper bound find $\bar{W}_V$ that satisfies

$$\psi(x/n, -\partial \bar{W}_V(x/n)) \leq 0.$$

- A function that satisfies this condition is

$$\bar{W}_V(x/n) = \langle \varrho, x/n \rangle - \log \rho_*^V,$$

  where $\varrho_i = \log \rho_i$ and $\rho_*^V = \max\{\rho_i : v_i = 1\}$.

- Build our splitting scheme out of this function, i.e. the importance function is given by $U(x/n) = \bar{W}_V(X/n)/\log(r)$.

Logarithmically Efficient Estimation of Overflow Probabilities

- Dean and Dupuis established that if we use the importance function U then the splitting estimator for $p_n^V(x)$ is logarithmically efficient, and number of particles created grows subexponentially in n.
- Similarly Dupuis and Wang (09) established that using subsolutions to PDE from previous slide you can construct logarithmically efficient IS estimators for overflow probabilities in Jackson networks.
- How do we then evaluate relative merits of two algorithms?
- Requires refined knowledge of performance characteristics, not just logarithmic scale.

Asymptotics of Overflow Probabilities in Jackson Networks

- Stationary distribution of Jackson networks:

$$\pi(m_1, \ldots, m_d) = \prod_{j=1}^{d} \mathbb{P}(Q_j(\infty) = m_j)$$
$$= \prod_{j=1}^{d} (1-\rho_j)\rho_j^{m_j}, \quad j = 1, \ldots, d, \text{ and } m_j \geq 0.$$

- Can use this result and a time-reversal argument to show that if x is in a compact set then there exists k_0 and k_1 such that

$$\limsup_{n\to\infty} \frac{p_n^v(x)}{e^{-n\gamma_v} n^{\beta_v - 1}} \leq k_1$$
$$\liminf_{n\to\infty} \frac{p_n^v(x)}{e^{-n\gamma_v} n^{\beta_v - 1}} \geq k_0$$

where where $\gamma_v = -\log \rho_*^v$, in which $\rho_*^v = \max\{\rho_i : v_i = 1\}$; and $\beta_v = \sum_i I\{\rho_i = \rho_*^v, v_i = 1\}$. See Blanchet (11), or Blanchet, Leder, Shi (11).

Computational Effort for Single Run of Splitting Algorithm

- In Blanchet, Leder, and Shi (11) we looked at the computational effort necessary to use a well designed splitting algorithm.
- Define $C = \frac{-\log \rho_*^v}{\log r}$, then rewrite importance function and level function as

$$U(x/n) = C(1 - \frac{\varrho}{\log \rho_*^v}\left(\frac{x}{n}\right))$$

$$\ell_n(x) = \lceil C(n - x\frac{\varrho}{\log \rho_*^v})\rceil.$$

- Consider the total number of particles that make it to overflow set, can see that

$$\mathbb{E}[N_n(x)] = r^{\ell_n(x)} p_n^V(x) \approx ce^{-\gamma_v n} n^{\beta_v - 1} r^{\ell_n(x)}.$$

- Notice that $e^{-\gamma_v} = e^{\log \rho_*^v} = e^{-C \log r} = r^{-C}$, so that

$$\mathbb{E}[N_n(x) \approx] cn^{\beta_v - 1} r^{\ell_n(x) - nC},$$

  and if we assume that $x/n \to 0$ then we have that $\mathbb{E}[N_n(x)] \approx cn^{\beta_v - 1}$

Refined Performance of Splitting

- From previous slide we saw that the number of particles to survive is $\approx n^{\beta_V - 1}$, the actual computational effort is on the order of $n^{\beta_V + 1}$ as established in Blanchet, Leder and Shi (11).
- The computational effort required to achieve a fixed level of relative error is given by

$$C_n \frac{\mathbb{E}[R_n(x)^2]}{p_n^V(x)^2},$$

  where C_n is the computational cost per replication of the estimator i.e. roughly $n^{\beta_V + 1}$.
- In Blanchet, Leder, Shi (11) we establish that $\mathbb{E}[R_n(x)^2] = p_n^V(x)^2 O\left(n^{\beta_V}\right)$.
- Thus the computational cost of a well designed splitting algorithm is $O\left(n^{2\beta_V + 1}\right)$.

Importance Sampling for Tandem Jackson Network

- Dupuis, Sezer, and Wang considered estimating total population overflow in d node tandem network using sampling measure defined as

$$\frac{\mathbb{Q}(Y(k) = z | Q(k-1) = x)}{\mathbb{Q}(Y(k) = z | Q(k-1) = x)} = \sum_{j=0}^{d} r_j(x/n) \exp\left(\langle \theta_j, z \rangle - \psi(\theta_j, x/n)\right),$$

where

$$r_j(x/n) = \frac{w_j(x/n)}{\sum_{j=0}^{2} w_k(x/n)}, \quad w_j(x/n) = \exp\left(n\langle \theta_j, x/n \rangle + n\gamma + j\delta n\right)$$

and

$$(\theta_j)_i = \begin{cases} \gamma, & 1 \leq i \leq d - j \\ 0, & \text{otherwise} \end{cases}$$

- Dupuis, Sezer and Wang established that this estimator is logarithmically efficient.
- Call associated estimator $\hat{p}_n$.

Refined Performance of Importance Sampling

- In Blanchet, Glynn and Leder (11) we performed refined analysis of this estimator to compare with splitting and other methods.
- We know that cost of algorithm is roughly

$$\frac{\mathbb{E}[\hat{p}_n]}{e^{-2n\gamma} n^{2\beta-2}}.$$

- By direct analysis of likelihood ratio on event of interest we are able to establish that

$$\mathbb{E}[\hat{p}_n] = O\left(e^{-2n\gamma} n^{2d}\right).$$

- We are able to establish that the computational complexity of this algorithm is $O\left(n^{2(d-\beta+1)}\right)$.

Comparing Performance on Estimating Overflow Probabilities in Tandem Networks

- Computational cost of splitting algorithm is $O(n^{2\beta+1})$.
- Computational cost of importance sampling algorithm is $O(n^{2(d-\beta+1)}$.
- Thus prefer importance sampling if more than half the stations are bottlenecks, and splitting otherwise.
- Conjecture: This property holds for all Jackson networks, not just tandem network topology.