Importance Sampling for Two Queuing Systems with Discontinuous Service Policies

Kevin Leder
Joint Work with Paul Dupuis and Hui Wang

September 24, 2008

1. Importance Sampling Review
2. Tandem-Queue with Server Slowdown
3. Serve the Longest Queue

Importance Sampling Basics

- Sequence of events $\{A_n\}$ in $(\Omega, \mathcal{F}, \mathbb{P})$ such that

$$\lim_n -\frac{1}{n} \log \mathbb{P}(A_n) = \gamma > 0.$$

- Estimate $\mathbb{P}(A_n)$ via IS by constructing sampling measure $\mathbb{Q}$ and averaging independent replications of

$$\hat{p}_n = 1_{A_n} \frac{d\mathbb{P}}{d\mathbb{Q}}.$$

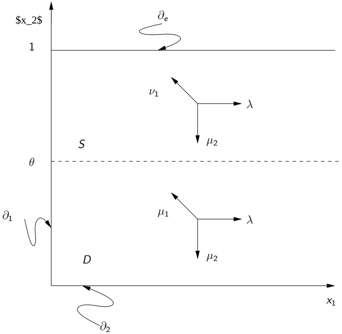
- Performance of estimator determined by $E^{\mathbb{Q}}[\hat{p}_n^2]$..
- Call a change of measure/estimator *asymptotically optimal* if

$$\liminf_n -\frac{1}{n} \log E^{\mathbb{Q}}[\hat{p}_n^2] \geq 2\gamma.$$

Tandem-Queue with Server Slowdown

State Space Description

Arrival rate λ. Service rate of first server is ν_1 or μ_1. Service rate of second server is μ_2. Q_i the size of queue i. For each $n \in \mathbb{N}$ $X_i^n = Q_i/n$ is the scaled process. Will focus on the case $\lambda < \nu_1 \leq \mu_2 \leq \mu_1$.



Problem Statement

- Goal is to develop *asymptotically optimal* importance sampling change of measure for estimating the buffer overflow probability

$$p_n = \mathbb{P}\left\{Q_2 = n \text{ before } Q = (0,0) \text{ starting from } (1,0)\right\}.$$

- Difficulty of this problem arises from discontinuity due to slowdown mechanism.
- Simple model of a queueing network with interior discontinuities.

Log Moment Generating Functions

1. Below the slow-down threshold [region D],

$$H(\alpha) \doteq \lambda(e^{\alpha_1} - 1) + \mu_1(e^{\alpha_2 - \alpha_1} - 1) + \mu_2(e^{-\alpha_2} - 1).$$

2. Above the slow-down threshold [region S],

$$H_s(\alpha) \doteq \lambda(e^{\alpha_1} - 1) + \nu_1(e^{\alpha_2 - \alpha_1} - 1) + \mu_2(e^{-\alpha_2} - 1).$$

3. On the boundary ∂_1,

$$H_{\partial_1}(\alpha) \doteq \lambda(e^{\alpha_1} - 1) + \mu_2(e^{-\alpha_2} - 1).$$

4. On the boundary ∂_2,

$$H_{\partial_2}(\alpha) \doteq \lambda(e^{\alpha_1} - 1) + \mu_1(e^{\alpha_2 - \alpha_1} - 1).$$

Local Rate Function

- X^n satisfies a large deviations principle (LDP) with local rate function $L(x, \beta)$.
- $L(\cdot, \cdot)$ built out of the Legendre transform of the log-moment generating functions from the previous slide: $L, L_s, L_{\partial_1}, L_{\partial_2}$ and their inf-convolutions.
- $L(x, \cdot)$ convex for each x, and $L(\cdot, \beta)$ sectionally homogenous.
- Important point is form of L on slowdown interface $x \in \{(x_1, \theta) : x_1 > 0\}$ then

$$\begin{aligned} L(x, \beta) &= [L \oplus L_s](\beta) \\ &= \inf \{\rho L(\beta_1) + (1-\rho) L_s(\beta_2) : \rho\beta_1 + (1-\rho)\beta_2 = \beta\} \\ &= \sup_{\alpha} [\langle \alpha, \beta \rangle - (H \vee H_s)(\alpha)] \end{aligned}$$

Exponential Decay Rate

- For the escape probability we have $\gamma \doteq -\lim \frac{1}{n} \log p_n$ where

$$\begin{aligned}\gamma &= \inf\left\{\int_0^T L(\phi,\dot{\phi})\,dt :\ \phi(0)=0, (\phi(T))_2 = 1, T \geq 0\right\} \\ &= \theta \log(\mu_2/\lambda) + (1-\theta)\log z\end{aligned}$$

- For any $x \in [0,\infty) \times [0,1]$ consider

$$\gamma(x) = \inf\left\{\int_0^T L(\phi,\dot{\phi})\,dt :\ \phi(0)=x, (\phi(T))_2 = 1, T \geq 0\right\}$$

Overview of PDE/Game Theoretic Approach to IS

- Fix scale parameter n and consider 2nd moment of a single sample under an IS change of measure.
- Interpret $-\frac{1}{n}\log[\text{2nd moment}]$ as a dynamic game. One player the IS change of measure–other player large deviation player.
- Consider the Isaacs equation obtained in the limit $n \to \infty$. If a function W satisfies the *subsolution* property, then it generates an IS change of measure, and a *verification argument* can be used to bound $-\frac{1}{n}\log[\text{2nd moment}]$ from below (here in terms of $W(0)$).
- The design problem: find W so $W(0)$ is big (here near 2γ) and so that the associated change of measure is easy to implement.

IS Change of Measure from Subsolution

- Suppose we are given a smooth subsolution W. Let $(p, q) = DW(x)$, and suppose we use new rates suggested by the saddle point in the game:
  1. for x such that $x_2 < \theta$ i.e. region D, use

$$\bar{r}^*(p, q) = (\bar{\lambda}, \bar{\mu}_1, \bar{\mu}_2), \text{ where } \bar{\lambda} = \lambda e^{-p/2}, \bar{\mu}_1 = \mu_1 e^{(p-q)/2}, \bar{\mu}_2 = \mu_2 e^{q/2},$$

  2. for x such that $x_2 \geq \theta$ i.e. region S, use

$$\bar{r}_s^*(p, q) = (\bar{\lambda}, \bar{\nu}_1, \bar{\mu}_2), \text{ where } \bar{\lambda} = \lambda e^{-p/2}, \bar{\nu}_1 = \nu_1 e^{(p-q)/2}, \bar{\mu}_2 = \mu_2 e^{q/2}.$$

- **Theorem.** Let W be a smooth subsolution and let V^n equal $-\frac{1}{n}\log[\text{2nd moment}]$ for a single sample under the associated scheme. Then

$$2\gamma \geq \limsup_n V^n \geq \liminf_n V^n \geq W(0).$$

Subsolutions to Isaacs Equation

- A *classical subsolution* to the Isaacs equation is a continuously differentiable function $\bar{W} : \mathbb{R}_+ \times [0,1] \to \mathbb{R}$ such that
  1. $\mathbb{H}(x, D\bar{W}(x)) \geq 0$ for all $x \in (0,\infty) \times (0,1)$,
  2. $\langle D\bar{W}(x), d_i \rangle \geq 0$ for all $x \in \partial_i$ where $d_i = -v_{i+1}$,
  3. $\bar{W}(x) \leq 0$ for all $x \in \partial_e$.

  where,
  1. $\mathbb{H}(x,\alpha) = -2H_s(-\alpha/2)$ if $x_2 \geq \theta$
  2. $\mathbb{H}(x,\alpha) = -2H(-\alpha/2)$ if $x_2 < \theta$.
- For smooth subsolution should also require $\mathbb{H} \wedge \mathbb{H}_s(D\bar{W}(x)) \geq 0$ if x is on interface.

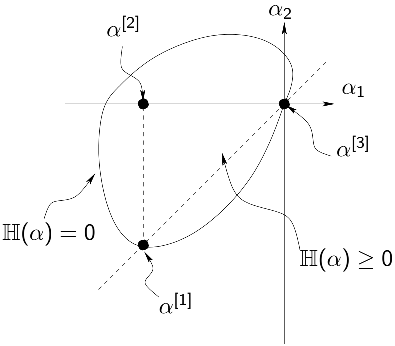
Construction of Subsolution

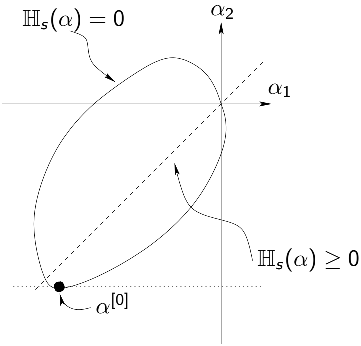
1. Construct collection of affine functions, these will make up the gradient of subsolution.
2. Define piecewise affine function as pointwise minimum of this collection.
3. Mollify piecewise affine function to get smooth function that satisfies subsolution property of Isaacs equation.

Important Roots

- Can be shown that $\mathbb{H}(x, 2D\gamma(x)) = 0$ and of course $\gamma(0) = \gamma$.
- By studying optimal exit trajectory can get idea of $D\gamma$ along optimal trajectory.
- From work of Miretskiy et. al. know form of optimal trajectory to be
  1. $(0,0) \to (0,\theta)$ vertical path pushing against left boundary.
  2. $(0,\theta) \to (0,1)$ vertical path gliding along left boundary.

Important Roots Cont'd





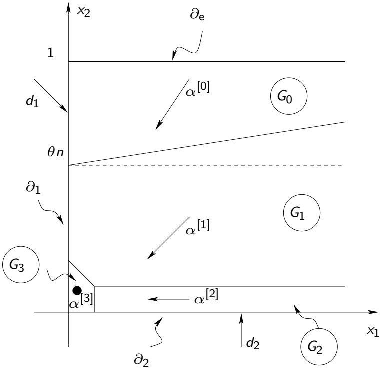
Construction of a Subsolution

- Fix an arbitrarily small $\delta > 0$ and define the affine functions

$$\begin{aligned}
\bar{W}_0^\delta(x) &= \langle x, \alpha^{[0]} \rangle + 2 \log z, \\
\bar{W}_1^\delta(x) &= \langle x, \alpha^{[1]} \rangle + 2\gamma - \delta, \\
\bar{W}_2^\delta(x) &= \langle x, \alpha^{[2]} \rangle + 2\gamma - 2\delta, \\
\bar{W}_3^\delta(x) &= \langle x, \alpha^{[3]} \rangle + 2\gamma - 3\delta.
\end{aligned}$$

- Define $\bar{W}^\delta \doteq \bar{W}_0^\delta \wedge \bar{W}_1^\delta \wedge \bar{W}_2^\delta \wedge \bar{W}_3^\delta$.
- Mollified version of W^δ is smooth subsolution that can be used for IS.

A Piecewise Affine Subsolution

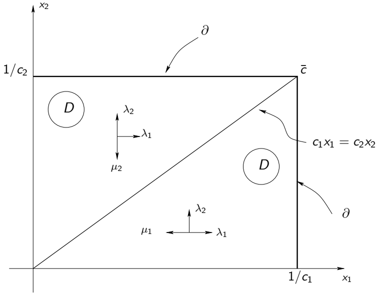


This figure depicts the gradients of $\bar{W}^{\delta}$ in different regions of the state space, with $\bar{W}^{\delta} = \bar{W}_i^{\delta}$ in region G_i.

Serve the Longest Queue

State space description

Weights c_i, arrival rates λ_i, and service rates $\mu_i, i = 1, 2$. Let Q_i be the size of queue i, and $X_i^n \doteq Q_i/n$ the scaled process. Priority to $\operatorname{argmax}\{c_1 Q_1, c_2 Q_2\}$. In case of tie priority goes to Q_2. Stability condition $\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} < 1$.



Problem Statement

- Construct asymptotically optimal importance sampling estimators for the following

$$p_n \doteq P\left\{X^n = (X_1^n, X_2^n) \in \partial \text{ before } X^n = 0\right\}$$

- Though stated in 2 dimensions results hold in higher dimensions.
- Difficulty due to discontinuity along line $c_1 Q_1 = c_2 Q_2$.

Large Deviations Rate Function

- For $i = 1, 2$ define

$$H^{(i)}(\alpha) \doteq \mu_i(e^{-\alpha_i} - 1) + \sum_{j=1}^{2} \lambda_j(e^{\alpha_j} - 1),$$

  and

$$L^{(i)}(\beta) = \sup_{\alpha \in \mathbb{R}^d} \left[\langle \alpha, \beta \rangle - H^{(i)}(\alpha) \right].$$

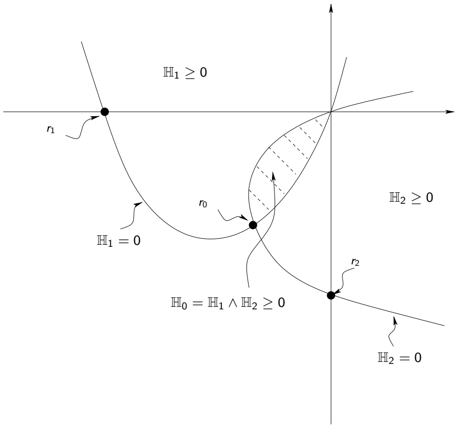
- For each $x \in \mathbb{R}_+^2$ define $\pi(x) = \{1 \leq i \leq 2 : c_i x_i = \max_j c_j x_j\}$, then define local rate function as

$$L(x, \beta) = \begin{cases} L^{(1)}(\beta), & \pi(x) = 1 \\ L^{(2)}(\beta), & \pi(x) = 2 \\ L^{(1)} \oplus L^{(2)}(\beta), & \pi(x) = 1, 2 \end{cases}$$

Isaacs Equation

- Build IS change of measure out of gradients of continuously differentiable functions $W : \mathbb{R}_+^2 \to \mathbb{R}$ that satisfy
  1. $\mathbb{H}(x, D\bar{W}(x)) \geq 0$ for all $x \in D$
  2. $\bar{W}(x) \leq 0$ for all $x \in \partial$.
- In the above $\mathbb{H}(x, \alpha) = \mathbb{H}_j(\alpha) \doteq -2H^{(j)}(-\alpha/2)$, where $j = \max\{i : i \in \pi(x)\}$.
- Only exit boundary conditions needed, unlike server slowdown model.
- Again for smooth subsolution should also require $\mathbb{H}_1 \wedge \mathbb{H}_2(D\bar{W}(x)) \geq 0$ if $c_1 x_1 = c_2 x_2$.

Important Roots



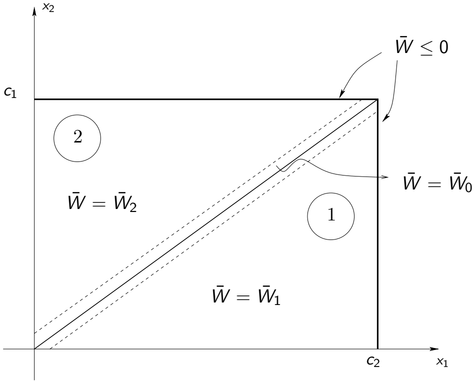
Construction of Smooth Subsolution

- For positive δ construct affine functions $\bar{W}_0, \bar{W}_1, \bar{W}_2$ out of roots r_0, r_1, r_2 such that pointwise minimum $\bar{W}$ satisfies
  1. $\bar{W}(0) = 2\gamma - \delta,$
  2. $\bar{W}(x) \leq 0$ for $x \in \partial$.
  3. $\bar{W}(x)$ has the form

$$\bar{W} = \begin{cases} \bar{W}_1(x), & \pi(x) = 1 \\ \bar{W}_2(x), & \pi(x) = 2 \\ \bar{W}_0(x), & \pi(x) = \{1, 2\} \end{cases}$$

- Use mollification technique to construct smooth subsolution, $W^{\varepsilon,\delta}$.

Form of Subsolution



This figure depicts the value of $\bar{W}$ in different regions of the state space.

Numerical Results

6 Dimension Example

	$n = 20$	$n = 50$	$n = 80$
Theoretical value	2.1×10^{-8}	4.2×10^{-13}	3.7×10^{-20}
Estimate	2.21×10^{-8}	4.21×10^{-13}	3.75×10^{-20}
Std. Err.	0.08×10^{-8}	0.15×10^{-13}	0.14×10^{-20}
95% C.I.	$[2.05, 2.37] \times 10^{-8}$	$[3.91, 4.50] \times 10^{-13}$	$[3.48, 4.02] \times 10^{-20}$

Table 3. $(\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6) = (1, 2, 2, 3, 1, 8), (\mu_1, \mu_2, \mu_3, \mu_4, \mu_5, \mu_6) = (10, 15, 16, 16, 15, 24)$, and $(c_1, c_2, c_3, c_4, c_5, c_6) = (1/2, 1, 1, 1, 1, 1/3)$

4 Dimension Example

	$n = 20$	$n = 50$	$n = 80$
Theoretical value	5.62×10^{-9}	1.54×10^{-14}	7.01×10^{-23}
Estimate	5.52×10^{-9}	1.51×10^{-14}	6.91×10^{-23}
Std. Err.	0.27×10^{-9}	0.09×10^{-14}	0.33×10^{-23}
95% C.I.	$[4.99, 6.04] \times 10^{-9}$	$[1.33, 1.69] \times 10^{-14}$	$[6.26, 7.56] \times 10^{-23}$

Table 2. $(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = (1, 2, 2, 4), (\mu_1, \mu_2, \mu_3, \mu_4) = (5, 12, 10, 15)$, and $(c_1, c_2, c_3, c_4) = (1/2, 1, 1, 1)$

Concluding Remarks

- Important property of both systems: form of local rate function along discontinuity i.e.

$$L(x,\beta) = L_1 \oplus L_2(\beta) \text{ and } L(x,\beta) = L \oplus L_s(\beta).$$

- In order to construct smooth subsolution in domain should use a gradient along interface that satisfies Isaacs equation on both sides of interface e.g.,

$$\mathbb{H}_1 \wedge \mathbb{H}_2(x) \geq 0 \text{ or } \mathbb{H} \wedge \mathbb{H}_s(D\bar{W}(x)) \geq 0.$$

- This condition agrees exactly with the above representation of the local rate function along interface.