State-dependent Large-deviations Based Importance Sampling for a Slow-down Tandem Queue

D.I. Miretskiy W.R.W. Scheinhardt M.R.H. Mandjes

July 14-18, 2008, Sandbjerg Estate

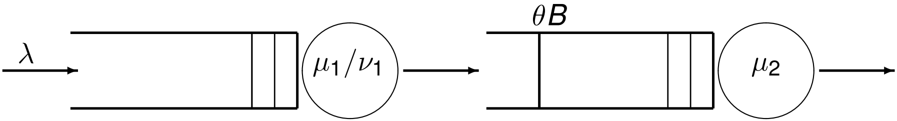
Efficient Monte Carlo: Conference on the Occasion of Rubinstein's 70th Birthday

Part of this research has been funded by the Dutch BSIK/BRICKS project

Outline

Outline

A Slow-down Tandem Queue



Arrival process:	*Poisson*(λ)
Service rates:	*exp*(μ_1), exp(ν_1), exp(μ_2)
Slow-down mechanism:	$\nu_1 < \mu_1$

Underlying processes

- Joint queue-length process

$$Q(t) = (Q_1(t), Q_2(t))$$

- Scaled queue-length process

$$X(t) = Q(Bt)/B$$

- Embedded (discrete-time) queue-length process

$$X_k = (X_{1,k}, X_{2,k})$$

Goal

Find probability of collecting B jobs in the second buffer before emptying the system, starting from any state

Formally

$$p_B^x := \mathbb{P}(X_2(T_B) = 1 | X(0) = x)$$

where

$$T_B := \min\{t > 0 | X_2(t) = 1 \text{ or } X(t) = 0\}$$

Goal

Find probability of collecting B jobs in the second buffer before emptying the system, starting from any state

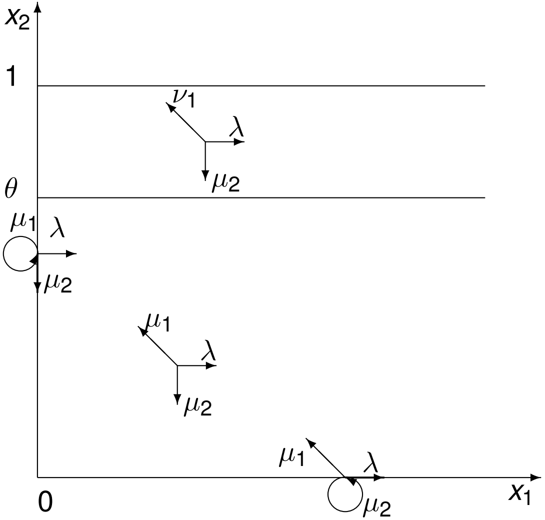
Formally

$$p_B^x := \mathbb{P}(X_2(T_B) = 1 | X(0) = x)$$

where

$$T_B := \min\{t > 0 | X_2(t) = 1 \text{ or } X(t) = 0\}$$

State Space



x_2
1
ν_1
λ
μ_2
θ
μ_1
λ
μ_2
μ_1
λ
μ_2
μ_1
λ
0
μ_2
x_1

Method

- State-dependent Importance Sampling
  - simulating the system under the new measure $\mathbb{Q}$, such that $\mathbb{Q}(X_2(T_B) = 1) \gg \mathbb{P}(X_2(T_B) = 1)$
  - the new measure is changing dynamically

- Asymptotic efficiency

$$\lim_{B \to \infty} \frac{\log \mathbb{E}(L^2 \mathbb{I})}{\log \mathbb{E}(L \mathbb{I})} \geq 2$$

Here, L is the likelihood ratio of the particular sample path; $\mathbb{I}(X_2(T_B) = 1) = 1$ and $\mathbb{I}(X(T_B) = 0) = 0$

Outline

Cost-functions

Following holds for any Poisson process $N(t)$ with intensity λ

$$\mathbb{P}(N(t) \approx \tilde{\lambda} t) \approx e^{I(\tilde{\lambda}|\lambda)t}$$

where $I(\tilde{\lambda}|\lambda)$ is the cost-function, given by

$$I(\tilde{\lambda}|\lambda) := \sup_{\theta} \left(\theta\tilde{\lambda} - \log \mathbb{E} e^{\theta N(1)} \right)$$

after some optimization

$$I(\tilde{\lambda}|\lambda) = \lambda - \tilde{\lambda} + \tilde{\lambda} \log \frac{\tilde{\lambda}}{\lambda}$$

Cost of the path

The cost of moving one unit up in the interior under the new measure is

$$\frac{I(\tilde{\lambda}|\lambda) + I(\tilde{\mu}_1|\mu_1) + I(\tilde{\mu}_2|\mu_2)}{\tilde{\mu}_1 - \tilde{\mu}_2} \qquad \text{if } x_2 < \theta \text{, and}$$

$$\frac{I(\tilde{\lambda}|\lambda) + I(\tilde{\nu}_1|\nu_1) + I(\tilde{\mu}_2|\mu_2)}{\tilde{\nu}_1 - \tilde{\mu}_2} \qquad \text{if } x_2 \geq \theta$$

$\gamma(x)$ is the minimal cost of the path $x \to (\cdot, 1)$

Large deviations result

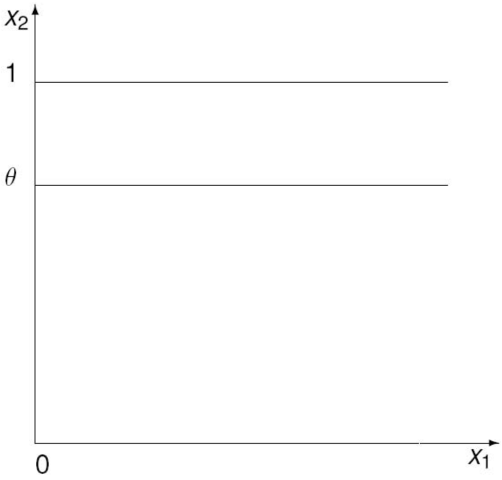
Theorem

$$\lim_{B \to \infty} \frac{1}{B} \log p_B^x = -\gamma(x)$$

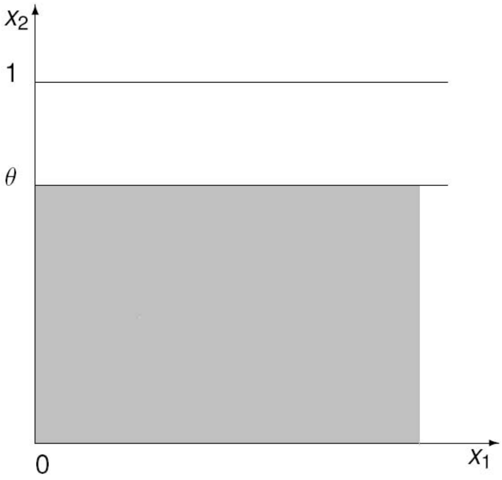
Remark

$\gamma(x)$ is the exponential decay rate of the interest probability.

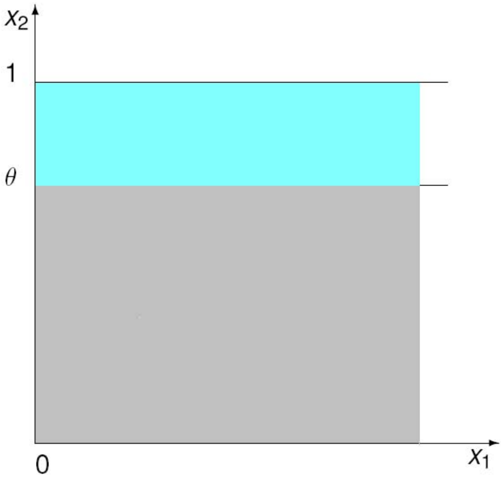
Five subsets with constant jump rates



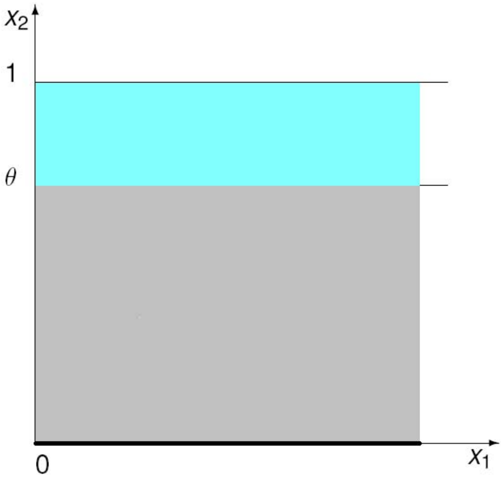
Five subsets with constant jump rates



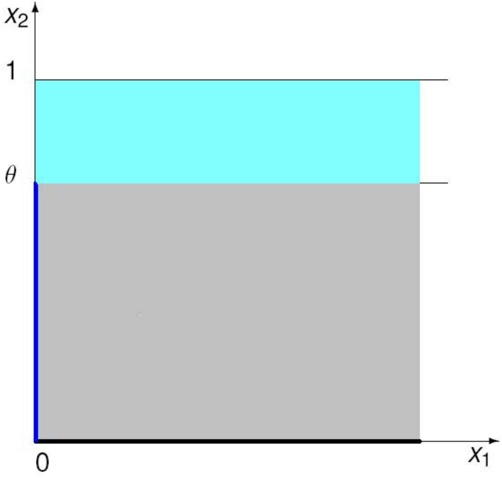
Five subsets with constant jump rates



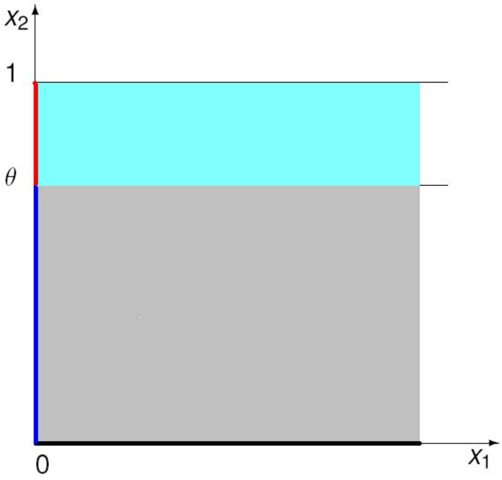
Five subsets with constant jump rates



Five subsets with constant jump rates



Five subsets with constant jump rates



The path structure

Lemma

The optimal path between any two states in the same subset is a straight line.

Lemma

The optimal path does not have more than

- *one subpath in each subset if* $\mu_2 < \mu_1$*, and*
- *two subpaths in each subset if* $\mu_2 \geq \mu_1$

The optimal path illumination

- Number of paths is limited (by Lemmas)
- Cost of each candidate is known
- So, we can find the shape of the optimal path to overflow in the second queue

The optimal path illumination

- Number of paths is limited (by Lemmas)
- Cost of each candidate is known
- So, we can find the shape of the optimal path to overflow in the second queue

The optimal path illumination

- Number of paths is limited (by Lemmas)
- Cost of each candidate is known
- So, we can find the shape of the optimal path to overflow in the second queue

Outline

Three possible cases

- $\mu_2 < \nu_1 < \mu_1$ i.e., the second server is the bottleneck
- $\nu_1 \leq \mu_2 < \mu_1$ i.e., the shifting bottleneck
- $\nu_1 < \mu_1 \leq \mu_2$ i.e., the first server is the bottleneck

We compute new jump rates

$$(\tilde{\lambda}(x), \tilde{\mu}_1(x), \tilde{\mu}_2(x)) \text{ and } (\tilde{\lambda}^+(x), \tilde{\nu}_1^+(x), \tilde{\mu}_2^+(x))$$

below and above the slow-down threshold

Three possible cases

- $\mu_2 < \nu_1 < \mu_1$ i.e., the second server is the bottleneck
- $\nu_1 \leq \mu_2 < \mu_1$ i.e., the shifting bottleneck
- $\nu_1 < \mu_1 \leq \mu_2$ i.e., the first server is the bottleneck

We compute new jump rates

$$(\tilde{\lambda}(x), \tilde{\mu}_1(x), \tilde{\mu}_2(x)) \text{ and } (\tilde{\lambda}^+(x), \tilde{\nu}_1^+(x), \tilde{\mu}_2^+(x))$$

below and above the slow-down threshold

Three possible cases

- $\mu_2 < \nu_1 < \mu_1$ i.e., the second server is the bottleneck
- $\nu_1 \leq \mu_2 < \mu_1$ i.e., the shifting bottleneck
- $\nu_1 < \mu_1 \leq \mu_2$ i.e., the first server is the bottleneck

We compute new jump rates

$$(\tilde{\lambda}(x), \tilde{\mu}_1(x), \tilde{\mu}_2(x)) \text{ and } (\tilde{\lambda}^+(x), \tilde{\nu}_1^+(x), \tilde{\mu}_2^+(x))$$

below and above the slow-down threshold

Three possible cases

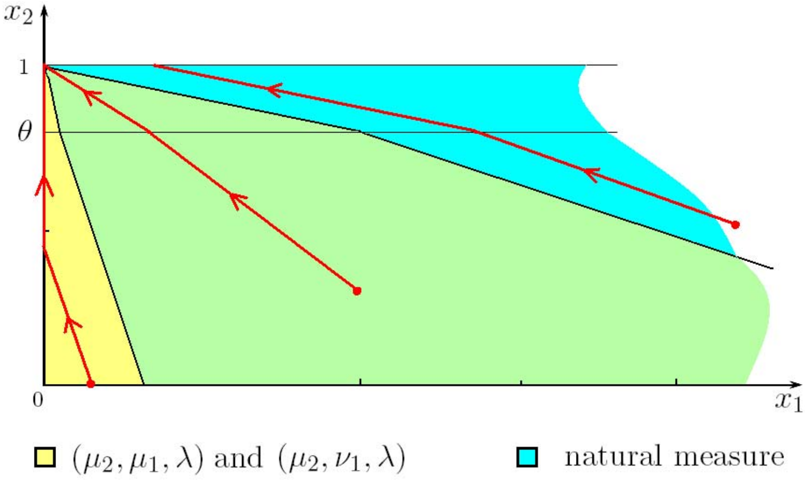
- $\mu_2 < \nu_1 < \mu_1$ i.e., the second server is the bottleneck
- $\nu_1 \leq \mu_2 < \mu_1$ i.e., the shifting bottleneck
- $\nu_1 < \mu_1 \leq \mu_2$ i.e., the first server is the bottleneck

We compute new jump rates

$$(\tilde{\lambda}(x), \tilde{\mu}_1(x), \tilde{\mu}_2(x)) \text{ and } (\tilde{\lambda}^+(x), \tilde{\nu}_1^+(x), \tilde{\mu}_2^+(x))$$

below and above the slow-down threshold

The second server is the permanent bottleneck

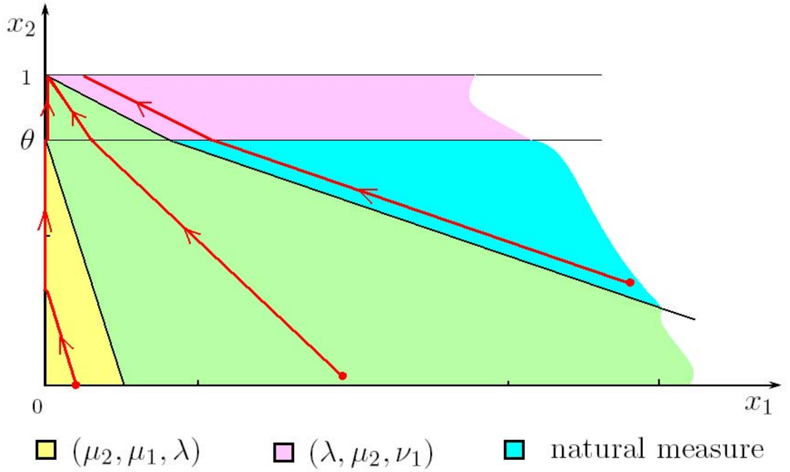


The green area

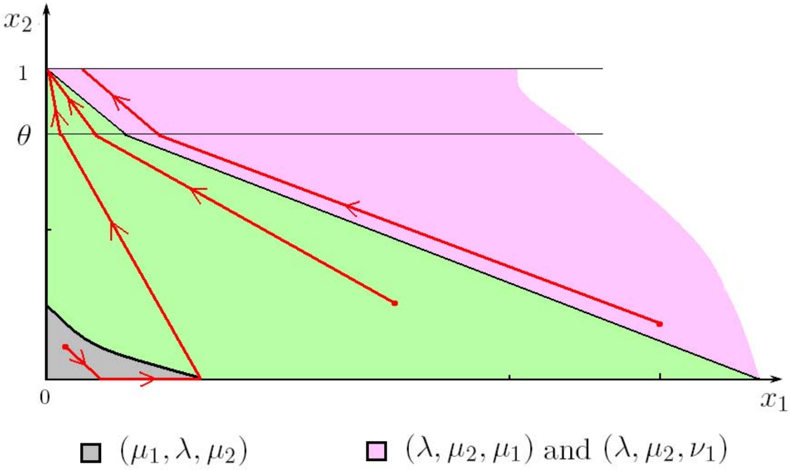
Solve

$$\begin{cases} \tilde\lambda = \tilde\mu_1 + \frac{\kappa^*(x)-x_1}{\theta-x_2}(\tilde\mu_1-\tilde\mu_2) \\ \tilde\lambda+\tilde\mu_1+\tilde\mu_2 = \lambda+\mu_1+\mu_2 \\ \tilde\lambda\tilde\mu_1\tilde\mu_2 = \lambda\mu_1\mu_2 \\ \\ \tilde\lambda^+ = \tilde\nu_1^+ - \frac{\kappa^*(x)}{1-\theta}(\tilde\nu_1^+-\tilde\mu_2^+) \\ \tilde\lambda^+ + \tilde\nu_1^+ + \tilde\mu_2^+ = \lambda+\nu_1+\mu_2 \\ \tilde\lambda^+\tilde\nu_1^+\tilde\mu_2^+ = \lambda\nu_1\mu_2 \\ \\ \tilde\lambda \le \tilde\mu_1, \quad \tilde\mu_1 > \tilde\mu_2, \quad \tilde\lambda^+ \le \tilde\nu_1^+ \text{ and } \tilde\nu_1^+ > \tilde\mu_2^+ \\ \tilde\lambda, \tilde\mu_1, \tilde\mu_2, \tilde\lambda^+, \tilde\nu_1^+, \tilde\mu_2^+ > 0 \\ \\ \kappa^*(x) := x_1 - \frac{\tilde\mu_1-\tilde\lambda}{\tilde\mu_1-\tilde\mu_2}(\theta-x_2) = y_1 + \frac{\tilde\nu_1^+-\tilde\lambda^+}{\tilde\nu_1^+-\tilde\mu_2^+}(y_2-\theta), \end{cases}$$

The shifting bottleneck

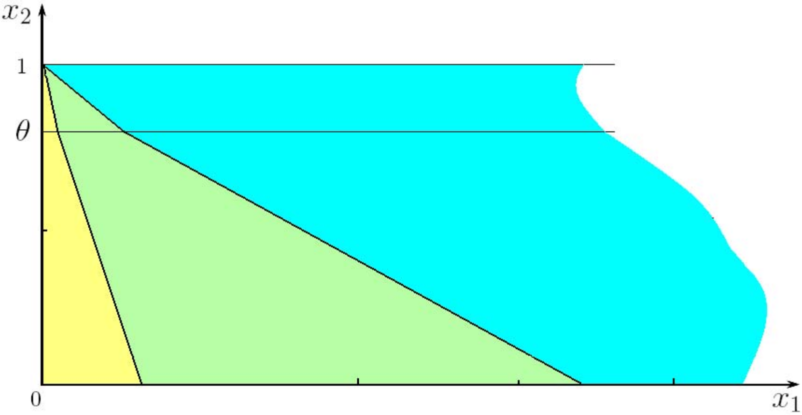


The first server is the permanent bottleneck

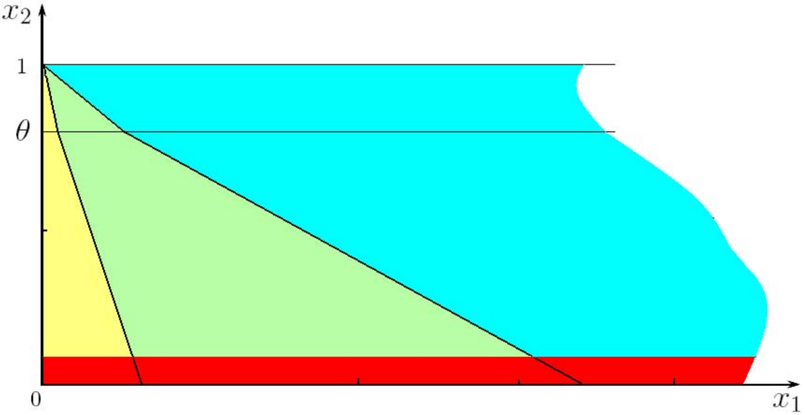


Outline

The protection line



The protection line



$W-$function

We define

$$
\begin{aligned}
W_1(x) &= 2\gamma_1(x_1, x_2) + 2\gamma_2(\kappa^*(x), \theta) - \delta && \text{if } x_2 < \theta, \\
W_1(x) &= 2\gamma_2(x_1, x_2) - \delta, && \text{if } x_2 \geq \theta \\
W_2(x) &= 2\gamma(x_1, \delta/2\gamma) - \delta, \\
W_3(x) &= 2\gamma - 3\delta,
\end{aligned}
$$

where

$$\gamma_1(x_1, x_2) := -(x_1 - \kappa^*(x)) \log \frac{\tilde{\lambda}(x_1, x_2)}{\lambda} - (\theta - x_2) \log \frac{\tilde{\mu}_2(x_1, x_2)}{\mu_2}$$

$$\gamma_2(\kappa^*(x), \theta) := -\kappa^*(x) \log \frac{\tilde{\lambda}^+(\kappa^*(x), \theta)}{\lambda} - (1 - \theta) \log \frac{\tilde{\mu}_2^+(\kappa^*(x), \theta)}{\mu_2},$$

$W-$function, mollification

$$W(x) = -\epsilon \log \sum_{i=1}^{3} e^{-W_i(x)/\epsilon}$$

Assumption

For positive parameters $\delta \equiv \delta_B$ and $\epsilon \equiv \epsilon_B$ the following holds:

$$\lim_{B\to\infty} \epsilon_B = 0, \qquad \lim_{B\to\infty} \delta_B = 0,$$

$$\lim_{B\to\infty} B\epsilon_B = \infty, \qquad \lim_{B\to\infty} \frac{\epsilon_B}{\delta_B} = 0$$

$W-$function, mollification

$$W(x) = -\epsilon \log \sum_{i=1}^{3} e^{-W_i(x)/\epsilon}$$

Assumption

For positive parameters $\delta \equiv \delta_B$ and $\epsilon \equiv \epsilon_B$ the following holds:

$$\lim_{B\to\infty} \epsilon_B = 0, \qquad \lim_{B\to\infty} \delta_B = 0,$$

$$\lim_{B\to\infty} B\epsilon_B = \infty, \qquad \lim_{B\to\infty} \frac{\epsilon_B}{\delta_B} = 0$$

The new measure

$$
\begin{aligned}
\bar{\lambda}(x) &= \lambda e^{-\langle DW(x), v_0\rangle/2} e^{\mathbb{H}(DW(x))/2}, && \text{if } x_2 < \theta \\
\bar{\mu}_i(x) &= \mu_i e^{-\langle DW(x), v_i\rangle/2} e^{\mathbb{H}(DW(x))/2}, && \text{if } x_2 < \theta, i = 1, 2, \\
\bar{\lambda}^+(x) &= \frac{\lambda}{\lambda+\nu_1+\mu_2} e^{-\langle DW(x), v_0\rangle/2} e^{\mathbb{H}^+(DW(x))/2}, && \text{if } x_2 \geq \theta \\
\bar{\nu}_1^+(x) &= \frac{\nu_1}{\lambda+\nu_1+\mu_2} e^{-\langle DW(x), v_1\rangle/2} e^{\mathbb{H}^+(DW(x))/2}, && \text{if } x_2 \geq \theta, \\
\bar{\mu}_2^+(x) &= \frac{\mu_2}{\lambda+\nu_1+\mu_2} e^{-\langle DW(x), v_2\rangle/2} e^{\mathbb{H}^+(DW(x))/2}, && \text{if } x_2 \geq \theta,
\end{aligned}
$$

Normalization

$$\mathbb{H}(DW(x)) =$$

$$2\log\left[\lambda e^{-\langle DW(x),v_0\rangle/2} + \mu_1 e^{-\langle DW(x),v_1\rangle/2} + \mu_2 e^{-\langle DW(x),v_2\rangle/2}\right]^{-1}$$

and

$$\mathbb{H}^{+}(DW(x)) =$$

$$2\log\left[\lambda e^{-\langle DW(x),v_0\rangle/2} + \nu_1 e^{-\langle DW(x),v_1\rangle/2} + \mu_2 e^{-\langle DW(x),v_2\rangle/2}\right]^{-1}$$

Outline

Likelihood ratio analysis

$$\begin{aligned}\log L(A) &= \frac{B}{2}\sum_{j=0}^{\sigma-1}\langle DW(X_j), X_{j+1}-X_j\rangle \\ &+ \sum_{k=1}^{2}\frac{1}{2}\sum_{j=0}^{\sigma-1}\langle DW(X_j), v_k\rangle I\{X_j = X_{j+1}, X_{k,j}=0\} \\ &- \frac{1}{2}\sum_{j=0}^{\sigma-1}\left(\mathbb{H}(DW(X_j))I_{\{X_{2,j}<\theta\}} + \mathbb{H}^+(DW(X_j))I_{\{X_{2,j}\geq\theta\}}\right).\end{aligned}$$

Minor terms

Lemma

For some positive constants γ^* *and* δ^*

$$\sum_{k=1}^{2} \frac{1}{2} \sum_{j=0}^{\sigma-1} \langle DW(X_j), v_k \rangle I\{X_j = X_{j+1} \in \partial_k\} \leq \gamma^* e^{-\delta^*/\epsilon} \sigma$$

Lemma

$$-\frac{1}{2}\left(\mathbb{H}(DW(X_j)) I_{\{X_j \in D\}} + \mathbb{H}^+(DW(X_j)) I_{\{X_j \in D^+\}}\right)$$
$$\leq -\tfrac{1}{2} e^{-(\theta - \frac{\delta}{2\gamma})\gamma/\epsilon} \mathbb{H}^+(DW_2(0, \theta))$$

Major term

Lemma

For any path $A = (X_j, j = 0, ..., \sigma)$, positive constants C and C^+

$$\frac{B}{2}\left|\sum_{j=0}^{\sigma-1}\langle DW(X_j), X_{j+1} - X_j\rangle - (W(X_\sigma) - W(X_0))\right| \leq \frac{C}{B\epsilon}\sigma + C^+\sigma^+,$$

where σ^+ is the number of the slow-down threshold crossings .

Lemma

For any υ_B such that $\lim_{B\to\infty}\upsilon_B = 0$ and T_B:

$$\lim_{B\to\infty}\frac{1}{B}\log\mathbb{E}(e^{\upsilon_B T_B}|X_2(T_B) = 1) = 0.$$

Major term

Lemma

For any path $A = (X_j, j = 0, ..., \sigma)$, positive constants C and C^+

$$\frac{B}{2}\left|\sum_{j=0}^{\sigma-1}\langle DW(X_j), X_{j+1} - X_j\rangle - (W(X_\sigma) - W(X_0))\right| \leq \frac{C}{B\epsilon}\sigma + C^+\sigma^+,$$

where σ^+ is the number of the slow-down threshold crossings .

Lemma

For any υ_B such that $\lim_{B\to\infty}\upsilon_B = 0$ and T_B:

$$\lim_{B\to\infty}\frac{1}{B}\log\mathbb{E}(e^{\upsilon_B T_B}|X_2(T_B) = 1) = 0.$$

Asymptotic efficiency

Theorem

$$\lim_{B\to\infty} \frac{1}{B} \log \mathbb{E}\left[L(A^x)\mathbb{I}(X_2(T_B)=1)\right] \le -2\gamma(x) = 2 \lim_{B\to\infty} \frac{1}{B} \log p_B^x,$$

The new measure

$(\bar{\lambda}(x), \bar{\mu}_1(x), \bar{\mu}_2(x))$ and $(\bar{\lambda}^+(x), \bar{\nu}_1^+(x), \bar{\mu}_2^+(x))$

is asymptotically efficient

Asymptotic efficiency

Theorem

$$\lim_{B\to\infty} \frac{1}{B} \log \mathbb{E}\left[L(A^x)\mathbb{I}(X_2(T_B)=1)\right] \leq -2\gamma(x) = 2\lim_{B\to\infty} \frac{1}{B} \log p_B^x,$$

The new measure

$(\bar{\lambda}(x), \bar{\mu}_1(x), \bar{\mu}_2(x))$ and $(\bar{\lambda}^+(x), \bar{\nu}_1^+(x), \bar{\mu}_2^+(x))$

is asymptotically efficient

Conclusions

- Our method provides very precise and accurate IS schemes
- The method is flexible and open for generalizing
- Slight modification of this method provides easy-to-implement and asymptotically efficient IS schemes

Conclusions

- Our method provides very precise and accurate IS schemes
- The method is flexible and open for generalizing
- Slight modification of this method provides easy-to-implement and asymptotically efficient IS schemes

Conclusions

- Our method provides very precise and accurate IS schemes
- The method is flexible and open for generalizing
- Slight modification of this method provides easy-to-implement and asymptotically efficient IS schemes