Easy and Efficient Importance Sampling Scheme for a Slow-down Tandem Queue

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

September 24-26, 2008, Rennes

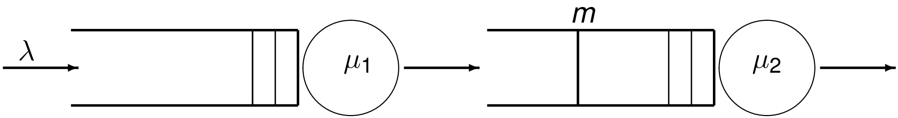
RESIM 2008

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

Outline

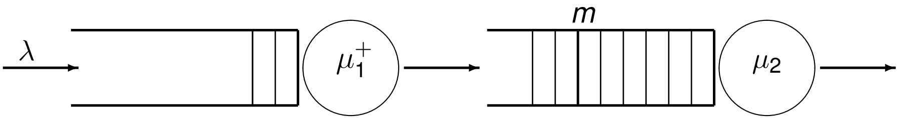
Outline

A Slow-down Tandem Queue



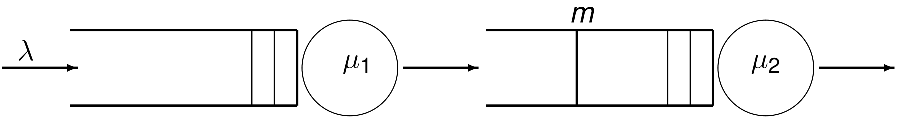
Arrival process:	*Poisson*(λ)
Service times:	*exp*(μ_1), exp(μ_1^+), exp(μ_2)
Slow-down mechanism:	$\mu_1^+ < \mu_1$

A Slow-down Tandem Queue



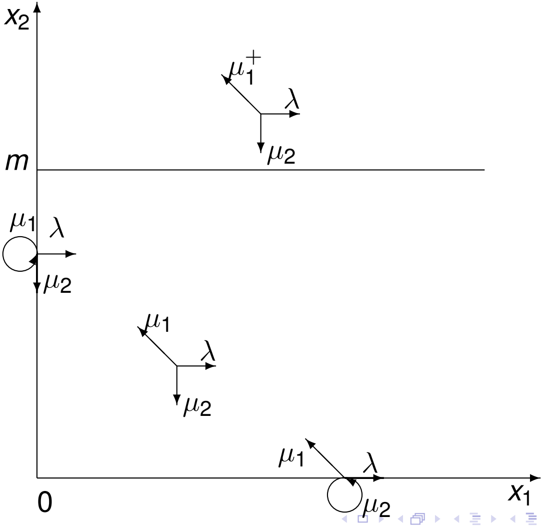
Arrival process:	*Poisson*(λ)
Service times:	*exp*(μ_1), exp(μ_1^+), exp(μ_2)
Slow-down mechanism:	$\mu_1^+ < \mu_1$

A Slow-down Tandem Queue



Arrival process:	*Poisson*(λ)
Service times:	*exp*(μ_1), exp(μ_1^+), exp(μ_2)
Slow-down mechanism:	$\mu_1^+ < \mu_1$

State Space


x_2
μ_1^+
λ
μ_2
m
μ_1
λ
μ_2
μ_1
λ
μ_2
μ_1
λ
0
μ_2
x_1

Stability of two-node Jackson tandem

Two-node Jackson tandem network with arrivals at rate λ, and service rates μ_1 and μ_2 is stable if, and only if

$$\lambda < \min\{\mu_1, \mu_2\}$$

- Slow-down network with $m = \infty$ is stable, iff

$$\lambda < \min\{\mu_1, \mu_2\}$$

- Slow-down network with $m = 0$ is stable, iff

$$\lambda < \min\{\mu_1^+, \mu_2\}$$

Stability of two-node Jackson tandem

Two-node Jackson tandem network with arrivals at rate λ, and service rates μ_1 and μ_2 is stable if, and only if

$$\lambda < \min\{\mu_1, \mu_2\}$$

- Slow-down network with $m = \infty$ is stable, iff

$$\lambda < \min\{\mu_1, \mu_2\}$$

- Slow-down network with $m = 0$ is stable, iff

$$\lambda < \min\{\mu_1^+, \mu_2\}$$

Stability of two-node Jackson tandem

Two-node Jackson tandem network with arrivals at rate λ, and service rates μ_1 and μ_2 is stable if, and only if

$$\lambda < \min\{\mu_1, \mu_2\}$$

- Slow-down network with $m = \infty$ is stable, iff

$$\lambda < \min\{\mu_1, \mu_2\}$$

- Slow-down network with $m = 0$ is stable, iff

$$\lambda < \min\{\mu_1^+, \mu_2\}$$

Stability of a slow-down system

Theorem

The slow-down system with parameters $(\lambda, \mu_1, \mu_1^+, \mu_2)$ *and slow-down threshold m is stable if, and only if*

$$\lambda < \frac{\mu_1|1-\psi^m||1-\psi^+| + \mu_1^+\psi^m|1-\psi|}{|1-\psi^m||1-\psi^+| + \psi^m|1-\psi|},$$

where $\psi = \mu_1/\mu_2$ *and* $\psi^+ = \mu_1^+/\mu_2$.

Outline

Underlying processes

- Joint queue-length process

$$Q_k = (Q_{1,k}, Q_{2,k})$$

- Scaled queue-length process

$$X_k = Q_k / B$$

Goal

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

Formally

$$p_B^s := \mathbb{P}(X_{2,T_B} = 1 | X_0 = x^s)$$

where

$$T_B := \min\{k > 0 | X_{2,k} = 1 \text{ or } X_k = 0\}$$

Goal

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

Formally

$$p_B^s := \mathbb{P}(X_{2,T_B} = 1 | X_0 = x^s)$$

where

$$T_B := \min\{k > 0 | X_{2,k} = 1 \text{ or } X_k = 0\}$$

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

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

Here, L is the likelihood ratio of the particular sample path; $\mathbb{I} = 1$, if $X_{2,T_B} = 1$ and 0 otherwise.

Large deviations result

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

Theorem

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

Remark

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

Large deviations result

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

Theorem

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

Remark

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

Large deviations result

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

Theorem

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

Remark

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

Result of minimization procedure

- The optimal trajectory
- The decay rate
- Pair of new measures

Outline

Three possible cases

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

We compute new jump rates

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

below and above the slow-down threshold

Three possible cases

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

We compute new jump rates

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

below and above the slow-down threshold

Three possible cases

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

We compute new jump rates

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

below and above the slow-down threshold

Three possible cases

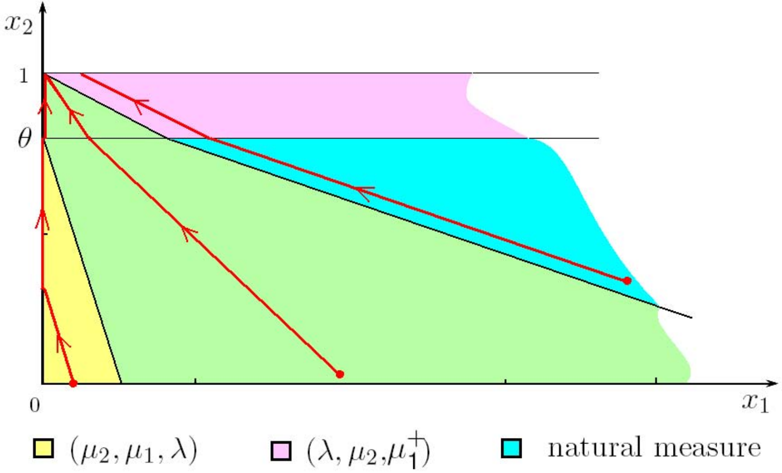
- $\mu_2 < \mu_1^+ < \mu_1$ i.e., the second server is the bottleneck
- $\mu_1^+ \leq \mu_2 < \mu_1$ i.e., the shifting bottleneck
- $\mu_1^+ < \mu_1 \leq \mu_2$ i.e., the first server is the bottleneck

We compute new jump rates

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

below and above the slow-down threshold

The shifting bottleneck



The green area

To solve after each transition

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

Efficiency

We were able to show that some modification of 'difficult' scheme is asymptotically efficient.

However, this method is very time consuming. So for practical aspects we need an easier tool.

'Easy' scheme...

Efficiency

We were able to show that some modification of 'difficult' scheme is asymptotically efficient.

However, this method is very time consuming. So for practical aspects we need an easier tool.

'Easy' scheme...

Efficiency

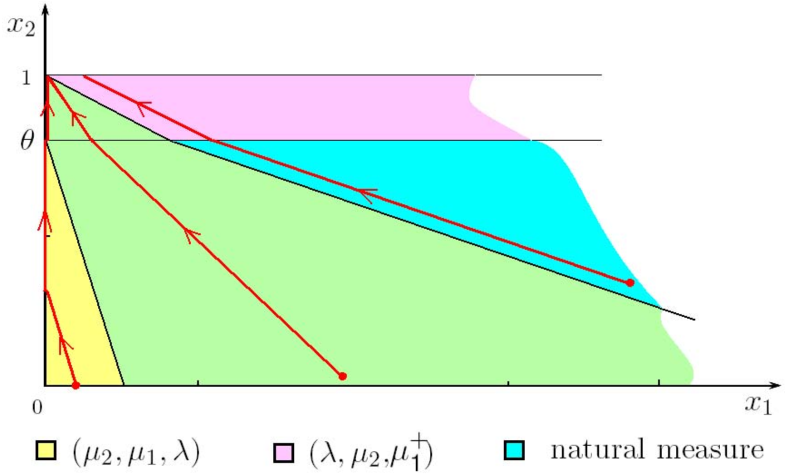
We were able to show that some modification of 'difficult' scheme is asymptotically efficient.

However, this method is very time consuming. So for practical aspects we need an easier tool.

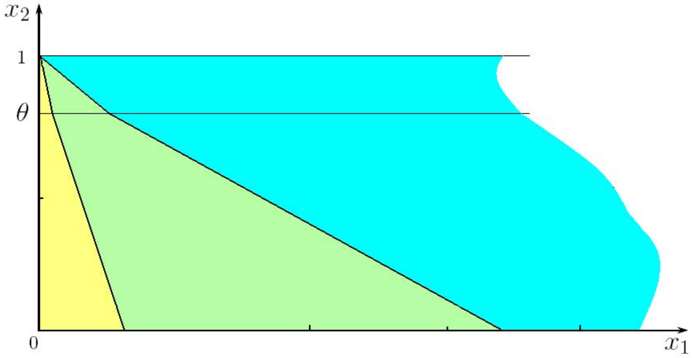
'Easy' scheme...

Outline

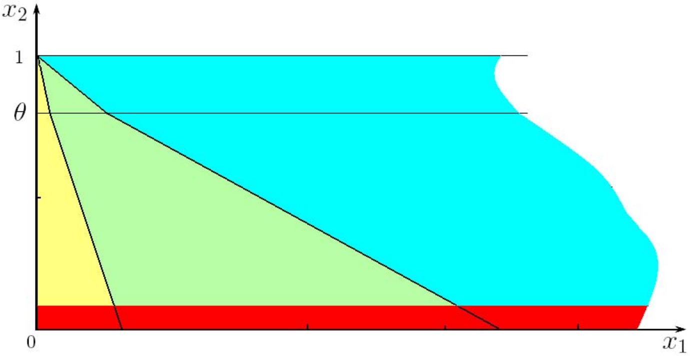
The shifting bottleneck



The protection line

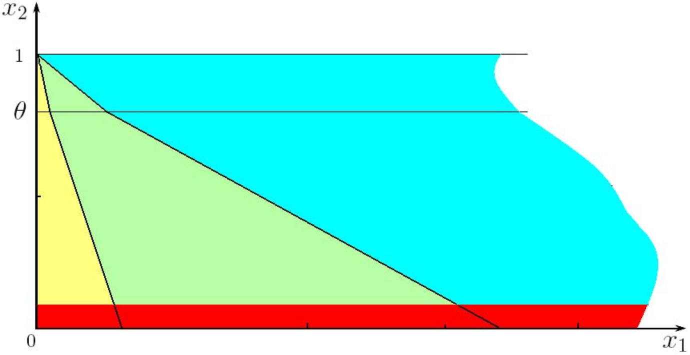


The protection line



$(\hat{\lambda}, \hat{\mu}_1, \hat{\mu}_2)$ and $(\hat{\lambda}^+, \hat{\mu}_1^+, \hat{\mu}_2^+)$

The protection line



$(\hat{\lambda}, \hat{\mu}_1, \hat{\mu}_2)$ and $(\hat{\lambda}^+, \hat{\mu}_1^+, \hat{\mu}_2^+)$

The new measure

$$
\begin{aligned}
\bar{\lambda}(x) &= (\tilde{\lambda})^{\rho_1(x)}(\hat{\lambda})^{\rho_2(x)}(\lambda)^{\rho_3(x)}M(x), \\
\bar{\mu}_1(x) &= (\tilde{\mu}_1)^{\rho_1(x)}(\hat{\mu}_1)^{\rho_2(x)}(\mu_1)^{\rho_3(x)}M(x) \qquad \text{if } x \in \bar{D}, \\
\bar{\mu}_2(x) &= (\tilde{\mu}_2)^{\rho_1(x)}(\hat{\mu}_2)^{\rho_2(x)}(\mu_2)^{\rho_3(x)}M(x),
\end{aligned}
$$

$$
\begin{aligned}
\bar{\lambda}^+(x) &= (\tilde{\lambda}^+)^{\rho_1(x)}(\hat{\lambda}^+)^{\rho_2(x)}(\lambda)^{\rho_3(x)}M^+(x), \\
\bar{\mu}_1^+(x) &= (\tilde{\mu}_1^+)^{\rho_1(x)}(\hat{\mu}_1^+)^{\rho_2(x)}(\mu_1^+)^{\rho_3(x)}M^+(x), \qquad \text{if } x \in \bar{D}^+, \\
\bar{\mu}_2^+(x) &= (\tilde{\mu}_2^+)^{\rho_1(x)}(\hat{\mu}_2^+)^{\rho_2(x)}(\mu_2)^{\rho_3(x)}M^+(x),
\end{aligned}
$$

$\gamma-$functions

$$\gamma(x) := \begin{cases} \gamma_1(x) + \gamma_2(\kappa, \theta) & \text{if } x \in \bar{D}, \\ \gamma_2(x) & \text{if } x \in \bar{D}^+. \end{cases}$$

with

$$\gamma_1(x) := -(x_1 - \kappa)\log\frac{\tilde{\lambda}}{\lambda} - (\theta - x_2)\log\frac{\tilde{\mu}_2}{\mu_2}, \quad \text{if } x \in \bar{D}$$

and

$$\gamma_2(x) := -x_1\log\frac{\tilde{\lambda}^+}{\lambda} - (1 - x_2)\log\frac{\tilde{\mu}_2^+}{\mu_2}, \quad \text{if } x \in \bar{D}^+$$

$\rho-$functions

We define

$$\rho_k(x) = \frac{e^{-W_k(x)/\epsilon}}{\sum_{i=1}^{3} e^{-W_i(x)/\epsilon}}, \quad k = 1, 2, 3$$

where

$$\begin{aligned} W_1(x) &= 2\gamma(x) - \delta, \\ W_2(x) &= 2\gamma(x_1, \delta/2 \log(\mu_2/\lambda)) - \delta, \\ W_3(x) &= 2\gamma(0) - 3\delta. \end{aligned}$$

Assumption

We choose positive parameters $\delta \equiv \delta_B$ and $\epsilon \equiv \epsilon_B$ such that (i) $\epsilon_B \to 0$, (ii) $\delta_B \to 0$, (iii) $B\epsilon_B \to \infty$, (iv) $\epsilon_B/\delta_B \to 0$, as $B \to \infty$.

$\rho-$functions

We define

$$\rho_k(x) = \frac{e^{-W_k(x)/\epsilon}}{\sum_{i=1}^{3} e^{-W_i(x)/\epsilon}}, \quad k = 1, 2, 3$$

where

$$\begin{aligned} W_1(x) &= 2\gamma(x) - \delta, \\ W_2(x) &= 2\gamma(x_1, \delta/2 \log(\mu_2/\lambda)) - \delta, \\ W_3(x) &= 2\gamma(0) - 3\delta. \end{aligned}$$

Assumption

We choose positive parameters $\delta \equiv \delta_B$ and $\epsilon \equiv \epsilon_B$ such that (i) $\epsilon_B \to 0$, (ii) $\delta_B \to 0$, (iii) $B\epsilon_B \to \infty$, (iv) $\epsilon_B/\delta_B \to 0$, as $B \to \infty$.

Outline

Likelihood ratio analysis

$$
\begin{aligned}
\log L(A) &= \frac{B}{2}\sum_{j=0}^{\sigma-1}\langle DW(X_j), X_{j+1}-X_j\rangle \\
&+ \sum_{k=1}^{2}\frac{1}{2}\sum_{j=0}^{\sigma-1}\langle DW(X_j), v_k\rangle I_{\{X_j=X_{j+1}\in\partial_k\}} \\
&- \sum_{j=0}^{\sigma-1}\left(\log M(x) I_{\{X_j\in D\}} + \log M^{+}(x) I_{\{X_j\in D^{+}\}}\right).
\end{aligned}
$$

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

For any $x \in D$ and for any $y \in D^+$ we have

$$\log M(x) \geq 0 \quad \textit{and} \quad \log M^+(y) \geq -C^\star e^{-h/\epsilon}$$

for some positive, finite constants $C^\star$ and h.

Major term

Lemma

For any path $A = (X_j, j = 0, ..., \sigma)$

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

where C *is some positive constant and* R *is the random error:*

$$R = \frac{B}{2} \log \left(\frac{\tilde{\lambda}}{\tilde{\lambda}^+} \right) \left| \sum_{i=1}^{\sigma^+} (-1)^i (\kappa - \eta_i) \right|,$$

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

Discontinuity error

Conjecture

For any scaled sample path $A = (X_j, j = 0, ..., \sigma)$ we believe the following holds true

$$\lim_{B\to\infty} \frac{1}{B} \log \mathbb{E}(e^R | X_{2,T_B} = 1) = 0.$$

$$R = \frac{B}{2} \log\left(\frac{\tilde{\lambda}}{\tilde{\lambda}^+}\right) \left|\sum_{i=1}^{\sigma^+} (-1)^i (\kappa - \eta_i)\right|.$$

Technical result

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

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

Outline

$(\lambda, \mu_1, \mu_1^+, \mu_2) = (0.1, 0.7, 0.15, 0.2)$, $\theta = 0.8$, $\epsilon = 0.001$

For relative error of 1%

x^s	B	p_B^s	# succ.	# runs	time
(0,0)	20	$3.79 \cdot 10^{-7} \pm 7.44 \cdot 10^{-9}$	15,576	28,332	8
	50	$1.28 \cdot 10^{-16} \pm 2.52 \cdot 10^{-18}$	33,542	58,332	45
	100	$3.54 \cdot 10^{-32} \pm 6.95 \cdot 10^{-34}$	56,982	109,992	163
(0.7B,0)	20	$6.12 \cdot 10^{-3} \pm 1.20 \cdot 10^{-4}$	8,946	8,946	1
	50	$3.73 \cdot 10^{-6} \pm 7.32 \cdot 10^{-8}$	24,665	24,665	9
	100	$3.28 \cdot 10^{-11} \pm 6.43 \cdot 10^{-13}$	51,365	51,365	36
(1.5B,0)	20	$5.18 \cdot 10^{-1} \pm 1.01 \cdot 10^{-2}$	11,287	12,888	< 1
	50	$1.35 \cdot 10^{-1} \pm 2.65 \cdot 10^{-3}$	66,997	77,942	2
	100	$1.05 \cdot 10^{-2} \pm 2.05 \cdot 10^{-4}$	316,351	367,327	21

Comparison of three different schemes

We fix # of replications at 10^6 and compare REs.

$(\lambda, \mu_1, \mu_1^+, \mu_2) = (0.1, 0.7, 0.15, 0.2)$			
B	st.-ind.	st.-dep.	current
20	$1.49 \cdot 10^{-3}$	$2.63 \cdot 10^{-3}$	$3.54 \cdot 10^{-3}$
50	$2.06 \cdot 10^{-3}$	$7.87 \cdot 10^{-3}$	$8.00 \cdot 10^{-3}$
100	$2.75 \cdot 10^{-3}$	$19.71 \cdot 10^{-3}$	$17.01 \cdot 10^{-3}$

$(\lambda, \mu_1, \mu_1^+, \mu_2) = (0.3, 0.36, 0.32, 0.34)$			
B	st.-ind.	st.-dep.	current
20	$0.92 \cdot 10^{-3}$	$5.30 \cdot 10^{-3}$	$6.00 \cdot 10^{-3}$
50	$12.50 \cdot 10^{-3}$	$8.40 \cdot 10^{-3}$	$11.00 \cdot 10^{-3}$
100	$39.69 \cdot 10^{-3}$	$12.20 \cdot 10^{-3}$	$11.00 \cdot 10^{-3}$

Conclusions

- The IS scheme is fast, easy and efficient
- Current IS scheme allows general starting states
- The method is flexible and can be applied generally

Conclusions

- The IS scheme is fast, easy and efficient
- Current IS scheme allows general starting states
- The method is flexible and can be applied generally

Conclusions

- The IS scheme is fast, easy and efficient
- Current IS scheme allows general starting states
- The method is flexible and can be applied generally