Importance Sampling for a Markov Modulated Queuing Network with Customer Impatience until the End of Service

Ebrahim MAHDIPOUR, Amir Masoud RAHMANI, Saeed SETAYESHI
IAU University, Tehran
mehdipour_msc@yahoo.com, rahmani@sr.iau.ac.ir, setayesh@aut.ac.ir

For more than two decades, there has been a growing of interest in fast simulation techniques for estimating probabilities of rare events in queuing networks. Importance sampling is a variance reduction method for simulating rare events. The present paper carries out strict deadlines to the paper by Dupuis et al for a two node tandem network with feedback whose arrival and service rates are modulated by an exogenous finite state Markov process. We derive a closed form solution for the probability of missing deadlines. Then we have employed the results to an importance sampling technique to estimate the probability of total population overflow which is a rare event. We have also shown that the probability of this rare event may be affected by various deadline values.
Keywords: *Importance Sampling, Queuing Network, Rare Event, Markov Process, Deadline*

1 Introduction

For more than two decades, there has been a growing of interest in fast simulation techniques for estimating probabilities of rare events in queuing networks. Among the available techniques importance sampling, a method in which the system is simulated under a different probability distribution (i.e., change of measure), has received much attention [3]. Importance sampling is a variance reduction method for simulating rare events. The idea in importance sampling is to change the sampling distribution (and modify the Monte Carlo estimator accordingly) to reduce estimator variance. The event of interest is total population overflow in a two-node Jackson network that allows feedback. In the other words given that initially the network is empty, the total number of customers in the network becomes *n* before the network empties [8]. Under the stability assumption and for large n one would expect this event to be rare.

In the present paper we develop an asymptotically optimal importance sampling technique for this network whose service and arrival processes are modulated by a finite state Markov chain. It is assumed that whenever a customer in node one has been served completely, it receives a deadline once entering the second queue, and it should finish its service and leave the system before missing this deadline. The difference between the deadline of a customer and its departure time from node one, referred to as a *relative deadline*, is a random variable η with a probability distribution function $D(\tau)$.

We consider a model with deterministic customer impatience. Customer service times and relative deadlines form sequences of iid random variables that are mutually independent. Given the number of customers in the system at any time, the future arrival process is assumed to be conditionally independent of the past history of the system.

In this paper, we have initially dealt with analysis and calculation of feedback rate in the network, and have shown its relation to relative deadline, arrival rate, and service rate too. Based on the feedback rate, we have also calculated the probability of missing deadline, and have used it for developing an asymptotically optimal importance sampling technique to estimate probability of total population overflow in Jackson network.

In [10] a similar sample of an asymptotically optimal importance sampling technique for estimating the probability of total population overflow is provided. In that system, however, the feedback probability is considered constant, and no deadline is assumed in the system.

In [11] a comprehensive study is given on the probability of missing the deadlines of

customers in M/M/1 queue. It supposes that the system drops a customer who misses its deadline. In the present study, we have extended the idea to a queue network. In addition, we do not drop a customer which misses its deadline, but we resend it to the first queue to be served again.

The paper is organized as follows. In Sections 2 basics of importance sampling and asymptotic optimality condition are described. Section 3 presents the structure and assumptions of the two-node Jackson network with feedback and its dynamics. A brief review of calculating the probability of missing the deadlines is discussed in section 4, and the proposed dynamic importance sampling algorithm is derived in Section 5. Section 6 examines examples to illustrate the efficacy of our method. Finally, section 7 concludes the paper.

2 Asymptotic Importance Sampling

We are interested in efficient importance sampling schemes for estimating the buffer overflow probability p_n when n is large. Importance sampling simulates the system under a different probability distribution, i.e., change of measure. Denote by A_n the event of buffer overflow, and rewrite $p_n = P(A_n)$. An importance sampling scheme generates samples from a new probability measure, say Q_n, such that $P << Q_n$. The estimator is then given by the average of independent replications of

$$\hat{p}_n \doteq 1_{A_n} \frac{dP}{dQ_n} \quad (2.1)$$

where dP/dQ_n is the Radon-Nikodym derivative [12] or likelihood ratio. Clearly $\hat{p}_n$ is unbiased for any such Q_n.

The goal of importance sampling is to choose Q_n to minimize the variance, or the second moment of $\hat{p}_n$. An obvious lower bound follows from Jensen's inequality [1] and the large deviations properties of $\hat{p}_n$,

$$\lim_n \inf \frac{1}{n} \log E^{Q_n}[\hat{p}_n^2] \geq \lim_n \inf \frac{2}{n} \log E^{Q_n}[\hat{p}_n]$$

$$= \lim_n \inf \frac{2}{n} \log p_n = -2\gamma \quad (2.2)$$

An importance sampling scheme, or the change of measure Q_n, is said to be *asymptotically optimal* if this lower bound is achieved, i.e., if

$$\lim_n \sup \frac{1}{n} \log E^{Q_n}[\hat{p}_n^2] \leq -2\gamma \quad (2.3)$$

For future analysis, it is worthwhile to note that the second moment equals

$$E^{Q_n}[\hat{p}_n^2] = E^P[\hat{p}_n] \quad (2.4)$$

3 Two-node Jackson network with feedback

Consider two-node Jackson network as in Fig. 1. The arrival and service rates of the system are determined by an exogenous Markov process [13] taking values in N={0, 1, 2, ..., n }, Let n be a positive integer.

The arrival process to the network is Poisson with rate λ and the arrival rate to the first queue is λ_1. Customers are served in the order of their arrival, i.e., service discipline is first-come-first-served (FCFS). Service times are exponentially distributed with rates μ_1 and μ_2 at node one and two. It is also assumed that whenever a customer in node one has been served completely, it receives a deadline once entering the second queue, and it should finish its service and leave the system before missing this deadline. In other words, the deadlines of customers in the second queue are effective until the end of their service at node two. If a customer misses its deadline irrespective of whether or not it is being served at node two, it must return to the first queue and wait again for receiving service in node one. The probability of missing deadline for customers in the second queue is φ_d and $1-\varphi_d$ is the probability of leaving the network without missing deadline. We assume that the two queues share one buffer with capacity n. Let $\bar{\mu} \doteq \mu_1 \wedge \mu_2$. Assuming the stability condition $\lambda < \bar{\mu}(1-\varphi_d)$, and without loss of generality, $\lambda + \mu_1 + \mu_2 = 1$, we have [14]

$$\gamma \doteq \lim_n -\frac{1}{n} \log p_n = \log \frac{(1-\varphi_d)\bar{\mu}}{\lambda} \quad (3.1)$$

The goal is to find an efficient importance sampling scheme for the estimation of p_n.

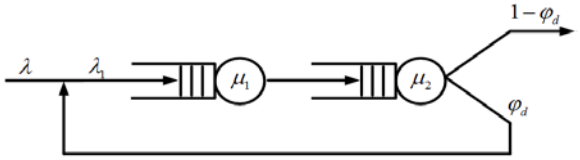

Fig. 1. Two-node network with feedback

A. System dynamics

Let N={N(k): k=0,1,2,…} be the embedded discrete time Markov chain that represents the queue lengths at the transition epochs of the network and suppose that $N(k)=(N_1(k),N_2(k))$ where $N_i(k)$ is the length of the queue at node i after the k^{th} transition. Then the dynamics of N can be modeled by N(k+1)=N(k)+π[N(k),Y(k+1)] where {Y(k)} are iid random variables taking values in:

$$\Omega=\{\omega_0=(1,0),\omega_1=(-1,1),\omega_2=(0,-1),\omega_3=(1,-1)\},$$

and the mapping π is defined as

$$\pi[N,y]\doteq\begin{cases}0, & \text{if } N_1=0 \quad \text{and } y=\omega_1\\ 0, & \text{if } N_2=0 \quad \text{and } y=\omega_2 \text{ or } \omega_3\\ y, & \text{otherwise}\end{cases}. \quad (3.2)$$

The distribution of N is completely determined by that of the sequence $Y=\{Y(k)\}$. Define $P^+(\Omega)\doteq\{\theta=(\theta_0,\theta_1,\theta_2)\}$ where θ is a probability on Ω and $\theta_i=\theta[\omega_i]$ for every $i=0,1,2$ under the original probability measure P, the distribution of $Y(k)$ is just

$$\Theta\doteq(\lambda,\mu_1,(1-\varphi_d)\mu_2,\varphi_d\mu_2)\in \mathrm{P}^+(\Omega) \quad (3.3)$$

See figure 2 for an illustration of the system dynamics.

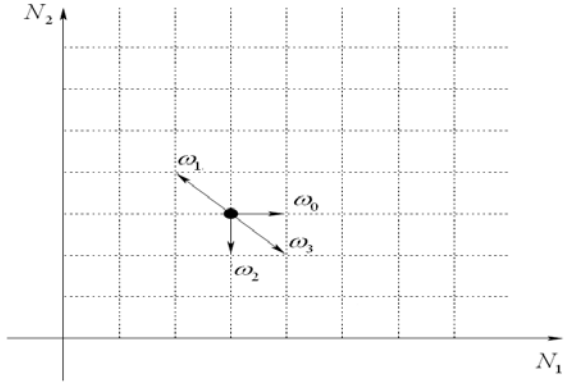

Fig. 2. State dynamics

4 Analysis of missing the deadlines

Each customer which departs node one and arrives at second queue receives a deadline, and it should be served in node two before missing its deadline. Thus, a customer who misses its deadline must return to the first queue immediately irrespective of whether or not it is being served. This type of customer behavior happens frequently in practice and has already been studied in references [2, 3].

The difference between the deadline of a customer in the second queue and its arrival time from node one, referred to as a *relative deadline*, is a random variable η known as customer impatience with a probability distribution function $D(\tau)$. In this paper, we consider a model with deterministic customer impatience. This type of customer impatience has been considered in references [5]. The probability distribution function of customer impatience η is given by

$$D(\bar{\eta},\tau)=0, \text{ if } \tau<\bar{\eta},$$
$$D(\bar{\eta},\tau)=1, \text{ if } \tau\geq\bar{\eta}, \quad (4.1)$$

where $\bar{\eta}$ is a constant denoting the mean customer impatience. Customer service times and relative deadlines form sequences of iid random variables that are mutually independent.

$V\equiv$ the time an arriving customer from node one with infinite (no) deadline must wait before it completes its service at node two in the long run. (4.2)

V is called the *offered sojourn time* in the system. The probability distribution function of *V is*

$$F_V(\tau)=P(V\leq\tau), \quad (4.3)$$

or, equivalently, the probability density function

$$f_V(\tau)=\frac{dF_V(\tau)}{d\tau} \quad (4.4)$$

Therefore the *probability of missing deadline*, defined as,

$$\varphi_d=P(\theta<V<\infty)=\int_0^\infty D(\tau)dF_V(\tau) \quad (4.5)$$

φ_d represents the steady-state probability that a customer misses its deadline. Let N denote the set of natural numbers (including 0) and $\mathbb{R}^+$ the set of positive real numbers. We also

use $E(m)$ to denote an Erlang random variable with parameters m and μ_2 ($E(0)=0$). Thus, $E(m)$ has the probability distribution function

$$F_{E(m)}(\mu_2,\tau)=P(E(m)\le\tau)$$
$$=1-e^{-\mu_2\tau}\sum_{i=0}^{m-1}\frac{(\mu_2\tau)^i}{i!},\ \text{if}\ m>0. \quad (4.6)$$

and

$\Psi_m(t,\varepsilon)\equiv$ the probability that one of the customers in the second queue during $[t,t+\varepsilon]$ misses its deadline, given there are m customers in the second queue at time t. (4.7)

$$\xi_m(t)=\lim_{\varepsilon\to 0}\frac{\Psi_m(t,\varepsilon)}{\varepsilon}, \quad (4.8)$$

$$\xi_m=\lim_{t\to\infty}\xi_m(t) \quad (4.9)$$

ξ_m is called the (*long-run*) *conditional loss rate* for m customers in the second queue. We derive a closed-form solution for the conditional probability density function of the offered sojourn time, given the number of customers in the second queue. More formally, let

$V_m\equiv$ the time an arriving customer from node one with infinite (no) deadline must wait in the second queue before it completes its service at node two in the long run, given it finds m customers in the second queue. (4.10)

V_m is called the *conditional offered sojourn time*, given there are m customers in the second queue. We now proceed to derive the probability density function of V_m. Let

$$F_{V_m}(\tau)=P(V_m\le\tau), \quad (4.11)$$

$$f_{V_m}(\tau)=\frac{dF_{V_m}(\tau)}{d\tau}, \quad (4.12)$$

Then we have

$$F_{V_0}(\tau)=1-e^{-\mu_2\tau},$$

$$F_{V_m}(\tau)=\frac{1}{P(V_{m-1}\le\eta_m)}\int_0^\tau(1-e^{-\mu_2(\tau-x)})(1-D(x))\ dF_{V_{m-1}}(x),$$

if $m\ge 1$, (4.13)

or, equivalently,

$$f_{V_0}(\tau)=\mu_2e^{-\mu_2\tau},$$

$$f_{V_m}(\tau)=\frac{\mu_2e^{-\mu_2\tau}}{P(V_{m-1}\le\eta_m)}\int_0^\tau f_{V_{m-1}}(x)e^{\mu_2x}(1-D(x))dx,\ \text{if}$$

$n\ge 1$, (4.14)

A solution for (4.14) may be given as

$$f_{V_0}(\tau)=\mu_2e^{-\mu_2\tau},$$

$$f_{V_m}(\tau)=\frac{\mu_2^{m+1}}{m!\prod_{k=1}^{m}P(V_{k-1}\le\eta_k)}\left[\int_0^\tau(1-D(x))dx\right]^m e^{-\mu_2\tau}$$

, if $m\ge 1$, (4.15)

Define $\Phi_m(s)$ to be the Laplace transform of $\left[\int_0^\tau(1-D(x))\right]^m$, i.e.,

$$\Phi_m(s)=\int_0^\infty\left[\int_0^\tau(1-D(x))\right]^m e^{-s\tau}d\tau. \quad (4.16)$$

Thus we have

$$f_{V_m}(\tau)=\frac{1}{\Phi_m(\mu_2)}\left[\int_0^\tau(1-D(x))dx\right]^m e^{-\mu_2\tau}. \quad (4.17)$$

$$\xi_0=0,$$

$$\xi_m=m\frac{\Phi_{m-1}(\mu_2)}{\Phi_m(\mu_2)}-\mu_2,\ \text{if}\ m>0. \quad (4.18)$$

For the case of deterministic customer impatience with $\bar{\eta}$ as the mean customer impatience $\Theta_m(s)$ can then be simplified as

$$\Phi_m(\bar{\eta},\mu_2)=\frac{m!}{\mu_2^{m+1}}F_{E(m)}(\mu_2,\bar{\eta}), \quad (4.19)$$

where $F_{E(m)}(\mu_2,\bar{\eta})$ is defined as in (4.6). Also, we find

$$\xi_0(\bar{\eta},\mu_2)=0,$$

$$\xi_m(\bar{\eta},\mu_2)=\mu_2\frac{F_{E(m-1)}(\mu_2,\bar{\eta})-F_{E(m)}(\mu_2,\bar{\eta})}{F_{E(m)}(\mu_2,\bar{\eta})},\ \text{if}$$

$m>0$, (4.20)

and the probability density function of the conditional offered sojourn time, $f_{V_m}(.)$, is given by

$$f_{V_m}(\tau)=\frac{\mu_2^{m+1}}{F_{E(m)}(\mu_2,\bar{\eta})m!}\tau^m e^{-\mu_2\tau},\quad \text{if}\ \tau<\bar{\eta},$$

$$f_{V_m}(\tau)=\frac{\mu_2^{m+1}}{F_{E(m)}(\mu_2,\bar{\eta})m!}\bar{\eta}^m e^{-\mu_2\tau},\quad \text{if}\ \tau\ge\bar{\eta},$$

(4.21)

It is clear that the maximum buffer size of the second queue equals n. Let $q_m(t)\equiv$ the probability that there are m customers in the system at time t. (4.22)

We can write

$$\frac{dq_0(t)}{dt} = -\lambda_1 q_0(t) + (\mu_2 + \xi_1(t))q_1(t),$$
$$\frac{dq_m(t)}{dt} = \lambda_1 q_{m-1}(t) - (\lambda_1 + \mu_2 + \xi_m(t))q_m(t)$$
$$+(\mu_2 + \xi_{m+1}(t))q_{m+1}(t), \text{ if } m \geq 1. \quad (4.23)$$

Where λ_1 is the arrival rate of customers to the first queue. Let $q_m = \lim_{t\to\infty} q_m(t)$ (4.24)

Then, in equilibrium, (4.23) is simplified as

$$0 = -\lambda_1 q_0 + (\mu_2 + \xi_1)q_1,$$
$$0 = \lambda_1 q_{m-1} - (\lambda_1 + \mu_2 + \xi_m)q_m + (\mu_2 + \xi_{m-1})q_{m+1}, \quad \text{if} \quad m \geq 1. \quad (4.25)$$

Using (4.18), we get

$$q_m = q_0 \frac{\mu_2 \lambda_1^m}{m!}\Phi_m(\mu_2), \text{ for } 1 \leq m \leq n, \quad (4.26)$$

The normalizing condition is

$$\sum_{i=0}^{n} q_i = 1, \quad (4.27)$$

which gives us

$$q_0 = \frac{1}{1 + \mu_2 \sum_{i=1}^{n} \lambda_1^i \frac{\Phi_i(\mu_2)}{i!}}. \quad (4.28)$$

Considering relations (4.18) and (4.26) we will have;

$$\xi = \sum_{i=0}^{n} q_i \xi_i \quad (4.29)$$

ξ represents loss rate of deadline of customers in second queue. It is clear that;

$$\lambda_1 = \lambda + \xi \quad (4.30)$$

Then we arrive at;

$$q_m = \frac{1}{1 + \mu_2 \sum_{i=1}^{n} \lambda_1^i \frac{\Phi_i(\mu_2)}{i!}} \times \frac{\mu_2 \lambda_1^m}{m!}\Phi_m(\mu_2),$$

and

$$\xi = \sum_{i=0}^{n} \left[\frac{1}{1 + \mu_2 \sum_{j=1}^{n} \lambda_1^j \frac{\Phi_j(\mu_2)}{j!}} \times \frac{\mu_2 \lambda_1^i}{i!}\Phi_i(\mu_2)\xi_i \right]$$

$$= \frac{\mu_2}{1 + \mu_2 \sum_{j=1}^{n} \lambda_1^j \frac{\Phi_j(\mu_2)}{j!}} \sum_{i=0}^{n} \frac{\lambda_1^i}{i!}\Phi_i(\mu_2)\ \xi_i \quad (4.31)$$

Considering the relation 4.30, we will have;

$$\lambda_1 = \lambda + \frac{\mu_2}{1 + \mu_2 \sum_{j=1}^{n} \lambda_1^j \frac{\Phi_j(\mu_2)}{j!}} \sum_{i=0}^{n} \frac{\lambda_1^i}{i!}\Phi_i(\mu_2)\ \xi_i$$

(4.32)

Relying on the numerical methods, we solve relation 4.32, and calculate λ_1 and ξ. The probability density function of customer offered sojourn time can then be determined as

$$f_V(\tau) = \sum_{i=0}^{n-1} q_i f_{V_i}(\tau), \quad (4.33)$$

The probability of missing deadline is derived as

$$\varphi_d = \sum_{i=0}^{n-1} q_i P(V_i > \eta_i)$$
$$= 1 - \sum_{i=0}^{n-1} q_i P(V_i \leq \eta_i)$$
$$= 1 - \sum_{i=0}^{n-1} \frac{q_i \mu_2 \Phi_{i+1}(\mu_2)}{(i+1)\ \Phi_i(\mu_2)} \quad (4.34)$$

η_i is a random variable with probability distribution function $D(.)$ denoting the relative deadline of the i^{th} customer in the second queue.

5 The dynamic importance sampling algorithm

The importance sampling schemes we consider use state-dependent changes of measure that can be characterized by stochastic kernels $\bar{\Theta}_n[.|.]$ on Ω given $\mathbb{Z}_+^2$, i.e, $\bar{\Theta}^n[.|x] \in P^+(\Omega)$ for every $x \in \mathbb{Z}_+^2$.

To be more precise, for a given threshold n, define the scaled state process $X_n = N/n$, where N is defined in section 3. Since the definition of π implies $\pi[n\,x, ny] = \pi[x, y]$ for every $x \in \mathbb{Z}_+^2$, it is not difficult to see that X_n satisfies the equation

$$X^n(k+1) = X^n(k) + \frac{1}{n}\pi[X^n(k), Y(k+1)] \quad (5.1)$$

with initial condition $X_n(0) = N(0)/n = 0$. The importance sampling generates $\{Y(k)\}$ as follows. The conditional probability of $Y(k+1) = \omega_i$, given $\{Y(j): j = 1, 2, ..., k\}$, is just $\bar{\Theta}_n[\omega_i \mid X_n(k)]$ for each i = 0, 1, 2. Define the hitting times

$$T_n \doteq \inf\{k \geq 0 : X_1^n(k) + X_2^n(k) = 1\}$$
$$T_0 \doteq \inf\{k \geq 0 : X_1^n(k) + X_2^n(k) = 0\}$$

Let A_n be the event of interest, that is,

$$A_n = \{X_1^n + X_2^n \quad reaches\, 1\, before\, returning\, to\, 0\}$$

$= \{T_n < T_0\}$

The importance sampling estimator is just

$$\hat{p}_n = 1_{A_n} \cdot \prod_{k=0}^{T_n-1} \frac{\Theta[Y(k+1)]}{\bar{\Theta}^n[Y(k+1) \mid X^n(k)]} \quad (5.2)$$

The second moment of $\hat{p}_n$, thanks to (2.4), equals $E^P[\hat{p}_n]$. The goal is to choose a stochastic kernel $\bar{\Theta}_n$ so that this second moment (whence the variance of $\hat{p}_n$) is as small as possible. Another important consideration is that one would like $\bar{\Theta}_n$ to be simple and easy to implement. Before we proceed to construct importance sampling algorithms, we collect some notation and terminology. Define

$\bar{D} \doteq \{(x_1, x_2) : x_i \geq 0, x_1 + x_2 \leq 1\}$

$D \doteq \{(x_1, x_2) : x_i \geq 0, x_1 + x_2 < 1\}$

$\delta_1 \doteq \{(0, x_2) : 0 < x_2 < 1\}$

$\delta_2 \doteq \{(x_1, 0) : 0 < x_1 < 1\}$

$\delta_e \doteq \{(x_1, x_2) : x_i \geq 0, x_1 + x_2 = 1\}$

$\bar{D}_n \doteq \bar{D} \cap \{(z_1, z_2)/n : (z_1, z_2) \in Z_+^2\}$

$D_n \doteq D \cap \{(z_1, z_2)/n : (z_1, z_2) \in Z_+^2\}$

Sometimes we refer to δ_e as the "exit boundary".

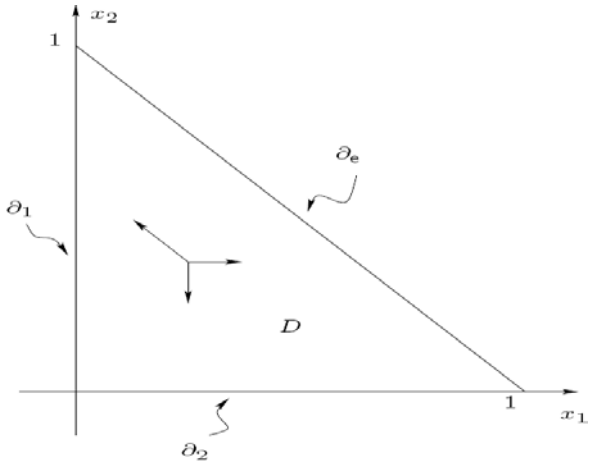

Fig. 3. Domains and boundaries

B. The Isaacs equation and boundary Hamiltonian

The main purpose of this section is to derive the Isaacs equation associated with the limit differential game that lies underneath importance sampling algorithms. The derivation will be kept formal. A rigorous argument, though possible, is not necessary for our purpose.

We can think of $\bar{\Theta}_n$ as a stochastic control problem and write down the corresponding Dynamic Programming Equation (DPE). To this end, we extend the dynamics and let, for every $x \in \bar{D}_n$,

$$V_n(x) \doteq \inf_{\bar{\Theta}_n} E_x^P[\hat{p}_n] = \inf_{\bar{\Theta}_n} E_x^P\left[1_{A_n} \cdot \prod_{k=0}^{T_n-1} \frac{\Theta[Y(k+1)]}{\bar{\Theta}^n[Y(k+1) \mid X^n(k)]}\right]$$

For simplicity, we further assume that $x \in D_n$, whence $\pi[x, y] \equiv y$ for every $y \in \Omega$. Under the original probability measure P, the sequence $\{Y(k)\}$ is iid with distribution Θ. Hence the DPE holds. Consider a logarithmic transform of V_n and define

$$W_n(x) \doteq -\frac{1}{n} \log V_n(x)$$

We have

$$W_n(x) = \sup_{\bar{\Theta} \in P^+(\Omega)} -\frac{1}{n} \log \sum_{i=0}^{3} \exp\left\{-nW_n\left(x + \frac{1}{n}\omega_i\right) - \log \frac{\bar{\Theta}[\omega_i]}{\Theta[\omega_i]}\right\} \Theta[\omega_i]$$

Remark 1. *Relative Entropy Representation for Exponential Integrals.*

Let (S, F) be a measurable space and $f : S \to R$ a bounded measurable function. Denote by $P(S)$ the space of probability measures on (S, F). Then for any $\gamma \in P(S)$,

$$-\log \int_S e^{-f} d\gamma = \inf_{\theta \in P(S)} \left[R(\theta \,||\, \gamma) + \int_S f \, d\theta\right]$$

Furthermore, the minimizer of the right-hand-side exists and is mutually absolutely continuous with respect to γ. Here the relative entropy $R(.\,||\,.)$ is defined as

$$R(\theta \,||\, \gamma) \doteq \begin{cases} \int_S \log \frac{d\theta}{d\gamma} d\theta, & \text{if } \theta << \gamma \\ \infty, & \text{otherwise} \end{cases}$$

A key step in the derivation is to apply the relative entropy representation for exponential integrals to the right-hand-side of the last equation. For completeness, we include the representation in its general form in Remark 1. It follows that

$$W_n(x) = \sup_{\bar{\Theta} \in P^+(\Omega)} \inf_{\theta \in P^+(\Omega)} \left[\sum_{i=0}^{3} W_n\left(x + \frac{1}{n}\omega_i\right) \theta[\omega_i] + \frac{1}{n}\left(\sum_{i=0}^{3} \theta[\omega_i] \log \frac{\bar{\Theta}[\omega_i]}{\Theta[\omega_i]} + R(\theta \,||\, \Theta)\right)\right]$$

Note that taking infimum over $\theta \in P^+(\Omega)$ is equivalent to taking infimum over $\theta \in P(\Omega)$

since by Remark 1 the minimizing θ is mutually absolutely continuous to Θ, whence it belongs to $P^+(\Omega)$. Suppose for now that $W_n(x)$ converges to $W(x)$. Formally assume the approximation

$$W_n\left(x+\frac{1}{n}\omega_i\right)- W_n(x) \approx \frac{1}{n}\langle DW(x),\omega_i\rangle$$

where DW is the gradient of W. Observing $\sum\theta[\omega_i] = 1$, we arrive at

$$0 = \sup_{\bar{\Theta}\in P^+(\Omega)}\inf_{\theta\in P^+(\Omega)}\left[\langle DW(x),F(\theta)\rangle + \sum_{i=0}^{3}\theta[\omega_i]\log\frac{\bar{\Theta}[\omega_i]}{\Theta[\omega_i]} + R(\theta\,\|\,\Theta)\right] \quad (5.3)$$

Where

$$F(\theta) \doteq \sum_{i=0}^{3}\theta[\omega_i].\omega_i \quad (5.4)$$

for each $\theta\in P^+(\Omega)$. Equation (3.6) is called an *Isaacs equation*. We now discuss the boundary conditions. For the exit boundary, we have by definition $V_n(x)=1$ or $W_n(x)=0$, therefore we impose the Dirichlet [9] boundary condition

$$W(x)=0, \text{ for } x\in\delta_e \quad (5.5)$$

For δ_1 and δ_2, we impose the Neumann [9] boundary condition that is typically associated with constrained dynamics

$$\langle DW(x),d_i\rangle = 0, \text{ for } x\in\delta_i \quad (5.6)$$

Finally, we make a few remarks on the game interpretation of importance sampling. The Isaacs equation (5.3) indicates that the underlying game has two players. The player who chooses the change of measure in order to minimize the second moment (i.e., $\bar{\Theta}$) becomes the maximizing player in the game due to the negative sign in the logarithmic transform. The minimizing player is artificially introduced, and chooses θ. We will refer to this player as the "large deviation player." The dynamics of the game are completely determined by θ, or the choice of the large deviation player, while the running cost of the game depends on the choices of both players.

Remark 2. The original dynamics have initial condition $x=0$, and $W(0)$ characterizes the asymptotic exponential decay rate of the second moment.

Following the argument (5.3), one can write down the Isaacs equation $H(DW(x))=0$ for $x\in D$, where

$$H(p) = \sup_{\bar{\Theta}\in P^+(\Omega)}\inf_{\theta\in P^+(\Omega)}\left[\langle p,F(\theta)\rangle + \sum_{i=0}^{3}\theta[\omega_i]\log\frac{\bar{\Theta}[\omega_i]}{\Theta[\omega_i]} + R(\theta\,\|\,\Theta)\right] \quad (5.7)$$

With the Dirichlet boundary condition $W(x)=0$ for $x\in\delta_e$.

However, as far as the boundaries δ_1 and δ_2 are concerned, the Neumann type boundary condition $\langle DW(x),d_i\rangle = 0$ is not sufficient (more precisely, it is not sufficient for δ_2, since the direction of constraint is not well defined on δ_2). Instead one has to resort to a *boundary Hamiltonian*, which, loosely speaking, is the Hamiltonian that one obtains using the state dynamics on the boundary [6]. Consequently, the boundary conditions become

$$H_{\delta_i}(DW(x)) = 0, \text{ for } x\in\delta_i, i=1,2, \quad (5.8)$$

where the boundary Hamiltonian H_{δ_i} is defined exactly as H except $F(\theta)$ is replaced by $F_i(\theta)$ with

$$F_1(\theta) = \sum_{i\neq j}\theta[\omega_i].\omega_i,$$
$$F_2(\theta) = \sum_{i\neq 2,3}\theta[\omega_i].\omega_i \quad (5.9)$$

Proposition 1. For every $p\in\mathbb{R}^d$, there exist a saddle point for the Hamiltonian H, say $(\bar{\Theta}^*(p),\theta^*(p))\in P^+(\Omega)\times P^+(\Omega)$, given by

$$\bar{\Theta}^*(p)[\omega_i] = \bar{\theta}^*(p)[\omega_i] = \Upsilon(p).\Theta[\omega_i]\exp\{-\langle p,\omega_i\rangle/2\}$$

where

$$\Upsilon(p) \doteq \left[\sum_{i=0}^{d}\Theta[\omega_i]\exp\{-\langle p,\omega_i\rangle/2\}\right]^{-1}$$

Moreover, the Hamiltonian H is concave and $H(p) = 2\log\Upsilon(p)$.

Remark 3. Proposition 1 can be easily applied to the interior Hamiltonian H and the boundary Hamiltonian H_{δ_i} to show the existence of saddle points and the concavity of these Hamiltonians. The formulae for the saddle points are as follows. Let $(\bar{\Theta}^*(.),\theta^*(.))$

be the saddle point for H, and $(\bar{\Theta}^*_{\delta_i}(.), \theta^*_{\delta_i}(.))$ be the saddle point for H_{δ_i}. Then

$$\bar{\Theta}^*(p) = \theta^*(p) = \Upsilon(p).\ \lambda(e^{-\frac{p_1}{2}}, \mu_1 e^{\frac{p_1-p_2}{2}}, (1-\varphi_d)\mu_2 e^{\frac{p_2}{2}}, \varphi_d \mu_2 e^{\frac{p_2-p_1}{2}})$$

$$\bar{\Theta}^*_{\delta_1}(p) = \theta^*_{\delta_1}(p) = \Upsilon_1(p).\ \lambda(e^{-\frac{p_1}{2}}, \mu_1, (1-\varphi_d)\mu_2 e^{\frac{p_2}{2}}, \varphi_d \mu_2 e^{\frac{p_2-p_1}{2}})$$

$$\bar{\Theta}^*_{\delta_2}(p) = \theta^*_{\delta_2}(p) = \Upsilon_2(p).\ \lambda(e^{-\frac{p_1}{2}}, \mu_1 e^{\frac{p_1-p_2}{2}}, (1-\varphi_d)\mu_2, \varphi_d \mu_2)$$

where $\Upsilon(p)$, $\Upsilon_i(p)$ are normalizing constants so that all these vectors are probability vectors (i.e., elements in $P^+(\Omega)$). Moreover, $H(p) = 2\log \lambda(p)$ and $H_{\delta_i}(p) = 2\log \lambda_i(p)$.

C. Piecewise affine subsolutions and mollification

The goal of this section is to construct piecewise affine subsolution $\bar{W}$ and its mollification. $\bar{W}$ must have the following properties (see Fig. 4).

1. The function $\bar{W}$ can be written as $\bar{W} = \bar{W}_1 \wedge \bar{W}_2 \wedge \bar{W}_3$ where $\bar{W}_k$ is an affine function for each *k* = 1, 2, 3.
2. $\bar{D}$ is divided into three regions R_1, R_2 and R_3, such that in each region R_k, $\bar{W} = \bar{W}_k$.
3. The subsolution property $H(D\bar{W}(x)) = H(D\bar{W}_k(x)) \geq 0$ holds for every x in the interior of region R_k.
4. The Dirichlet boundary inequality $\bar{W}(x) \leq 0$ for $x \in \delta_e$.
5. The Neumann boundary inequality $\langle D\bar{W}(x), d_i \rangle \geq 0$, whenever $x \in \delta_i$ and $D\bar{W}(x)$ exists.

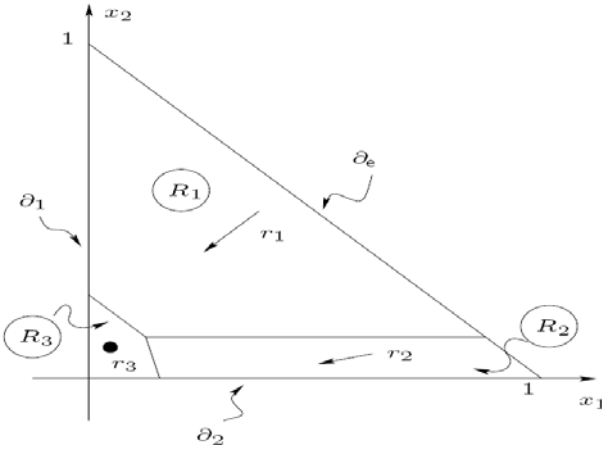

Fig. 4. The piecewise affine subsolution

This can be easily achieved – indeed, fixing an arbitrary $\delta > 0$, one can let, for each k,

$$\bar{W}_k^{\delta}(x) \doteq \langle r_k, x \rangle + 2\gamma - k\delta \quad (5.10)$$

where the r_k are depicted in Fig. 5. It is not difficult to check that satisfies all the requirements for all small $\delta > 0$.

Define the regions $R_i \subset \mathbb{R}^2$ $i \in \{0,1,2\}$:

$$R_1 \doteq \left\{x \in \mathbb{R}^2 : \bar{W}_1^{\delta}(x) \leq \bar{W}_2^{\delta}(x) \wedge \bar{W}_3^{\delta}(x)\right\}$$

$$R_2 \doteq \left\{x \in \mathbb{R}^2 : \bar{W}_2^{\delta}(x) \leq \bar{W}_1^{\delta}(x) \wedge \bar{W}_3^{\delta}(x)\right\}$$

$$R_3 \doteq \left\{x \in \mathbb{R}^2 : \bar{W}_3^{\delta}(x) \leq \bar{W}_1^{\delta}(x) \wedge \bar{W}_2^{\delta}(x)\right\}$$

These regions are depicted in Fig. 4. Note that

$$\bar{W}^{\delta}(x) = \bar{W}_i^{\delta}(x) \text{ for } x \in R_i$$

$$\bar{W}^{\delta} \doteq \bar{W}_1^{\delta} \wedge \bar{W}_2^{\delta} \wedge \bar{W}_3^{\delta}$$

satisfies all the requirements for all small $\delta > 0$. Define

$$r_1 \doteq 2\gamma(-1,-1)$$

$$r_2 \doteq 2\gamma(-1,0) + 2(\gamma - \alpha)(0,-1)$$

$$r_3 \doteq (0,0)$$

where α is given by

$$\alpha \doteq \begin{cases} \log\left[\mu_1/(\mu_1 + \lambda - (1-\beta)\mu_2)\right], & \text{if } \mu_1 \geq \mu_2 \\ \log\left[\mu_1/(\lambda + \beta\mu_1)\right], & \text{if } \mu_1 < \mu_2 \end{cases}$$

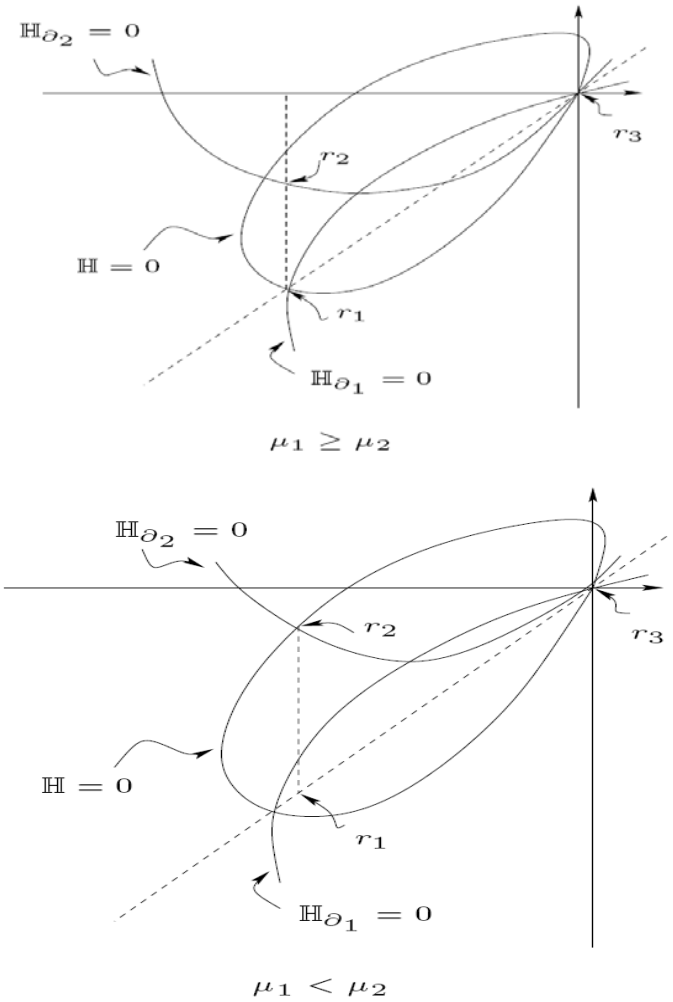

Fig. 5. The Hamiltonians and the choice of $\{r_k\}$

It is not difficult to check that $0 < \alpha \leq \gamma$. According to equation 5.10 we have

$\bar{W}_1^{\delta}(x) \doteq \langle r_1, x \rangle + 2\gamma - \delta$

$\bar{W}_2^{\delta}(x) \doteq \langle r_2, x \rangle + 2\gamma - 2\delta$

$\bar{W}_3^{\delta}(x) \doteq \langle r_3, x \rangle + 2\gamma - (1 + 2\gamma/\alpha)\delta$

The exponential weighting of $\bar{W}^{\delta}$ with parameter ε yields a smooth function

$$W^{\varepsilon,\delta}(x) \doteq -\varepsilon \log \sum_{k=1}^{3} \exp\left\{-\frac{1}{\varepsilon}\bar{W}_k^{\delta}(x)\right\}$$

that satisfies

$$DW^{\varepsilon,\delta}(x) = \sum_{k=1}^{3} \rho_k^{\varepsilon,\delta}(x) r_k$$

$$\rho_i^{\varepsilon,\delta}(x) \doteq \frac{\exp\left\{-\bar{W}_i^{\delta}(x)/\varepsilon\right\}}{\sum_{k=1}^{3} \exp\left\{-\bar{W}_k^{\delta}(x)/\varepsilon\right\}}$$

We have the following result.

Lemma 1. For each k we have $H(r_k) \geq 0$, and the function $W^{\varepsilon,\delta}$ satisfies

1. $H(DW^{\varepsilon,\delta}(x)) \geq 0$ for $x \in D$,
2. $W^{\varepsilon,\delta}(x) \leq 0$ for $x \in \delta_e$,
3. for each $i = 1, 2,$ and $x \in \delta_i$,

$$H_{\delta_i}(DW^{\varepsilon,\delta}(x)) \geq \sum_{k=1}^{3} \rho_k^{\varepsilon,\delta}(x) H_{\delta_i}(r_k)$$

$$\geq -\bar{C} \exp\{-\delta/\varepsilon\}$$

for some constant $\bar{C}$ that only depends on the system parameter Θ.

D. The importance sampling scheme and its asymptotics

The importance sampling scheme based on $\bar{W}^{\varepsilon,\delta}$ is as follows. Define the stochastic kernel $\bar{\Theta}^{\varepsilon,\delta}[.|.]$ on Ω by

$$\bar{\Theta}^{\varepsilon,\delta}[.|x] \doteq \sum_{k=1}^{3} \rho_k^{\varepsilon,\delta} \bar{\Theta}^{*}(r_k), \text{ if } x \in D \quad (5.11)$$

and

$$\bar{\Theta}^{\varepsilon,\delta}[.|x] \doteq \sum_{k=1}^{3} \rho_k^{\varepsilon,\delta} \bar{\Theta}^*_{\delta_i}(r_k), \text{ if } x \in \delta_i \quad (5.12)$$

Here the formulae for $\bar{\Theta}^*$ and $\bar{\Theta}^*_{\delta_i}$ can be found in Remark 3.

We will allow ε and δ to be *n*-dependent, denoted by ε_n, δ_n and let $\bar{\Theta}^n[.|.] \equiv \bar{\Theta}^{\varepsilon_n,\delta_n}[.|.]$.

Theorem 1. The importance sampling estimator $\hat{p}_n$ is asymptotically optimal if $\delta_n \to 0$, $\varepsilon_n/\delta_n \to 0$, and $n\varepsilon_n \to \infty$.

One can also use a fixed pair of parameters ε and δ for all n, According to [7] a good choice may be to take $\delta_n = -\varepsilon_n \log \varepsilon_n$.

6 Numerical results

Two experiments have been investigated to show the relation between the relative deadline, η, and the probability of missing deadline, φ_d then we use the results in the importance sampling technique to estimate the probability of total population overflow in the network. In the first experiment $\mu_2 = 0.5$, and for various values of λ, it is shown that how a change in η affects the probability of missing deadline. As illustrated in Fig. 6, the increase in relative deadline reduces φ_d like a decay function. Further increase in the relative deadline moves the probability of missing deadline towards zero, in such case, the possibility of missing deadline of customers is very low. On the contrary, assigning low values for η increases the probability of missing customers' deadline leading to raising the feedback rate in the network. In this experiment for $\lambda = 0.1$ we have also studied the relation between η and φ_d for some values of μ_2. As shown in Fig. 7, corresponding to a definite measure of the relative deadline, the increase in μ_2 decreases the probability of missing deadline, as the result of increase in the customers' service rate at node two, and decrease of their waiting time at the second queue.

The customers' arrival rate at the second queue is λ_1, thanks to the stability assumption in the network, thus the value of μ_1 will not affect the probability of missing deadline.

In the second experiment we evaluate the performance of the network for arbitrary values of the offered load ($\frac{\mu_2}{\lambda}$). Fig. 8 represents the probability of missing deadline for a set of offered loads, $n = 12$ and various values of η. In this figure the network has the best performance for $\eta = 6$ and it has the worst one for $\eta = 1$.

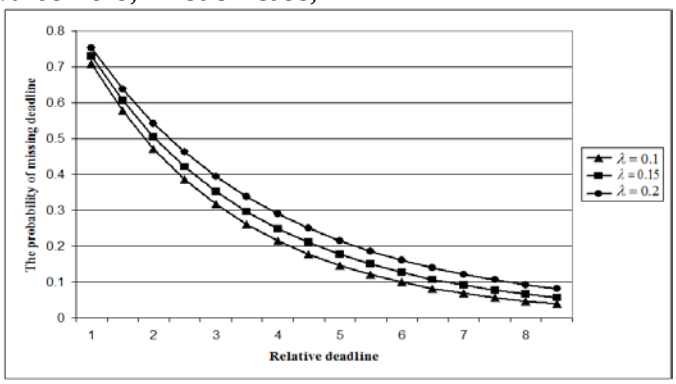

Fig. 6. Probability of missing deadline for various arrival rates

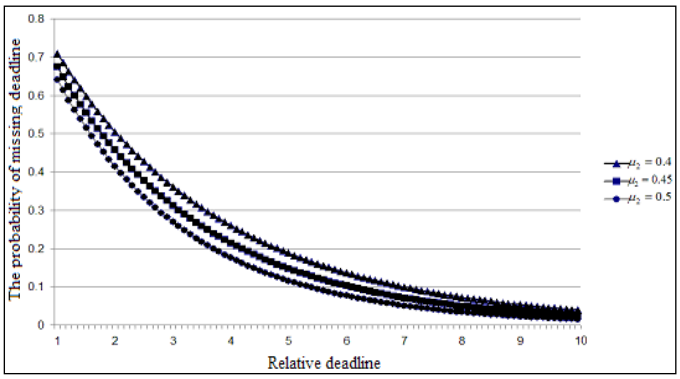

Fig. 7. Probability of missing deadline for various service rates

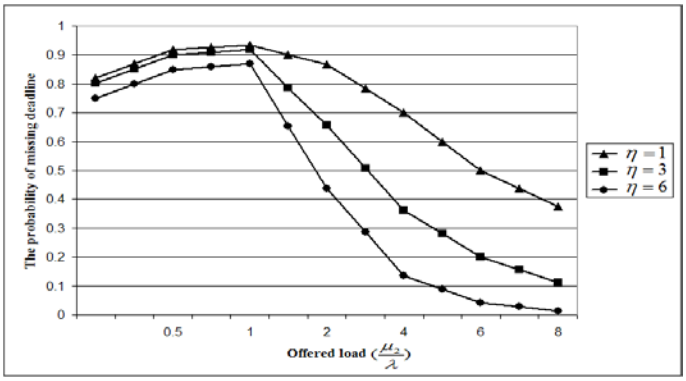

Fig. 8. Probability of missing deadline for various service rates

The above results have been used in the importance sampling technique to estimate the probability of total population overflow in the network for arbitrary values of φ_d. For the buffer size of the network we set n=12, and each estimate consists of 15000 replications. We run simulations for $\varphi_d = 0.2$, 0.4 and 0.6.

Table 1. Importance sampling estimation, $\lambda = 0.1$, $\mu_1 = 0.5$, $\mu_2 = 0.4$

	$\varphi_d = 0.2$	$\varphi_d = 0.4$	$\varphi_d = 0.6$
Theoretical value	0.82×10^{-5}	0.34×10^{-3}	0.34×10^{-1}
Estimate	0.79×10^{-5}	0.22×10^{-3}	0.36×10^{-1}
Std. Err	0.05×10^{-5}	0.02×10^{-3}	0.01×10^{-1}
95% C.I.	$[0.68,0.89]\times10^{-5}$	$[0.21,0.39]\times10^{-3}$	$[0.33,0.42]\times10^{-1}$

Table 2. Importance sampling estimation, $\lambda = 0.1$, $\mu_1 = 0.4$, $\mu_2 = 0.5$

	$\varphi_d = 0.2$	$\varphi_d = 0.4$	$\varphi_d = 0.6$
Theoretical value	0.31×10^{-5}	0.12×10^{-3}	0.19×10^{-1}
Estimate	0.39×10^{-5}	0.14×10^{-3}	0.16×10^{-1}
Std. Err	0.05×10^{-5}	0.01×10^{-3}	0.02×10^{-1}
95% C.I.	$[0.28,0.49]\times10^{-5}$	$[0.11,0.21]\times10^{-3}$	$[0.14,0.21]\times10^{-1}$

In Table 1 we set $\lambda = 0.1$, $\mu_1 = 0.5$ and $\mu_2 = 0.4$, In Table 2 we set $\lambda = 0.1$, $\mu_1 = 0.4$ and $\mu_2 = 0.5$. In Tables 3 and 4 the previous simulations repeated for n=20.

In all tables, "Std. Err" stands for "Standard Error" and "C.I." for "Confidence Interval". The results of simulations indicate that the performance of the importance sampling technique is stable across different

simulations, with estimates that are close to the theoretical value with small standard errors.

Table 3. Importance sampling estimation, $\lambda = 0.1$, $\mu_1 = 0.5$, $\mu_2 = 0.4$

	$\varphi_d = 0.2$	$\varphi_d = 0.4$	$\varphi_d = 0.6$
Theoretical value	1.81×10^{-9}	9.21×10^{-7}	7.95×10^{-3}
Estimate	1.74×10^{-9}	9.82×10^{-7}	7.66×10^{-3}
Std. Err	0.06×10^{-9}	0.53×10^{-7}	0.20×10^{-3}
95% C.I.	$[1.61, 1.87] \times 10^{-9}$	$[8.79, 10.86] \times 10^{-7}$	$[7.26, 8.05] \times 10^{-3}$

Table 4. Importance sampling estimation, $\lambda = 0.1$, $\mu_1 = 0.4$, $\mu_2 = 0.5$

	$\varphi_d = 0.2$	$\varphi_d = 0.4$	$\varphi_d = 0.6$
Theoretical value	1.17×10^{-9}	3.47×10^{-7}	1.77×10^{-3}
Estimate	1.20×10^{-9}	3.93×10^{-7}	1.72×10^{-3}
Std. Err	0.22×10^{-9}	0.32×10^{-7}	0.04×10^{-3}
95% C.I.	$[0.77, 1.63] \times 10^{-9}$	$[3.30, 4.57] \times 10^{-7}$	$[1.65, 1.80] \times 10^{-3}$

7 Conclusion

In this paper, we deal with the concept of deadline on a two-node tandem queue with feedback in which arrival and service rates are modulated by an exogenous finite state Markov process. Under such condition, we have extended importance sampling technique of Dupuis et. al. They have considered constant feedback rate with no connection to offered load of the network. Based on the definition of deadline for customers in the second queue, we have calculated the probability of missing deadline, and have shown how the feedback rate of the network is affected by the deadline value. We applied the feedback rate in importance sampling technique for estimating the probability of total buffer overflows. Our proposed method achieves more reality than previous works.

References

[1] P. Dupuis, A. Sezer and H. Wang, "Dynamic importance sampling for queueing networks," *The Annals of Applied Probability.*, vol. 17, no. 3, pp. 1306-1346, Jan. 2007.

[2] S. Asmussen, *Applied Probability and Queues*. New York: Spring, 2003, ch.4.

[3] S. Juneja and P. Shahabuddin, *Handbook on Simulation*, Amsterdam: Elsevier, 2006, ch.11.

[4] C. S. Chang, S. Juneja and P. Shahabuddin, "Effective bandwidth and fast simulation of ATM intree networks," *Performance Evaluation.*, vol. 20, no. 1, pp. 45-65, May. 1994.

[5] A. Movaghar, "On queueing with customer impatience until the end of service," *Stochastic Models.*, vol. 22, no. 1, pp. 149-173, May. 2006.

[6] D. Y. Barrer, "Queuing with Impatient Customers and Ordered Service," *Operations Research.*, vol. 5, no. 5, pp. 650-656, Oct. 1957.

[7] P. Dupuis and H. Wang. Subsolutions of an Isaacs equation and efficient schemes for importance sampling: Convergence analysis. *Preprint*, 2005.

[8] P. Dupuis and H. Wang. Subsolutions of an Isaacs equation and efficient schemes for importance sampling: Examples and numerics. *Preprint*, 2005.

[9] L. Kleinrock, Queueing Systems, Volume 1: Theory. New York: John Wiley & Sons, 1975, pp. 417.

[10] P. T. De Boer and V. F. Nicola, "Adaptive state-dependent importance sampling simulation of markovian queueing networks," *European Transactions on Telecommunications.*, vol.13, pp.303–315, Apr. 2002.

[11] A. Movaghar, "On queuing with customer impatience until the end of

service," in *Proc. 4th IEEE International Computer Performance and Dependability Symposium*, pp. 167–174, Chicago, 2000.

[12] P. T. De Boer, "Analysis of state-independent importance sampling measures for the two-node tandem queue," *ACM Trans. Modeling Comp. Simulation,* vol.16, pp.225–250, Jul. 2006.

Ebrahim MAHDIPOUR received his B.S. in computer engineering from IAU University, Dezful, in 2004, the M.S. in computer engineering from IAU University, Tehran, in 2006 and he is now pursuing the PhD degree in computer engineering at the IAU University, Tehran. His research interests are in the areas of queuing networks, computer networks, modeling and performance evaluation.

Amir Masoud RAHMANI received his B.S. in computer engineering from Amir Kabir University, Tehran, in 1996, the M.S. in computer engineering from Sharif University of technology, Tehran, in 1998 and the PhD degree in computer engineering from IAU University, Tehran, in 2005. He is assistant professor in the Department of Computer and Mechatronics Engineering at the IAU University. He is the author/co-author of more than 80 publications in technical journals and conferences. He served on the program committees of several national and international conferences. His research interests are in the areas of distributed systems, ad hoc and sensor wireless networks, scheduling algorithms and evolutionary computing.

Saeed SETAYESHI received his PhD on Electrical Engineering in Canada in 1993. His research interests are in the areas of Intelligent Control (Neural – Fuzzy – Expert), Adaptive Signal Processing, Nuclear Reactor Adaptive Controlling, Optical Electronics, Earthquake Prediction, LCD Design Knowledge, Base System, Data Analysis. He is assistant professor in the Department of Nuclear Engineering and Physics at the IAU University