Combining Importance Sampling and Temporal Difference Control Variates To Simulate Markov Chains

R. S. RANDHAWA
Stanford University
and
S. JUNEJA
Tata Institute of Fundamental Research

It is well known that in estimating performance measures associated with a stochastic system a good importance sampling distribution (IS) can give orders of magnitude of variance reduction while a bad one may lead to large, even infinite, variance. In this paper we study how this sensitivity of the estimator variance to the importance sampling change of measure may be "dampened" by combining importance sampling with stochastic approximation based temporal difference (TD) method. We consider a finite state space discrete time Markov chain (DTMC) with one-step transition rewards and an absorbing set of states and focus on estimating the cumulative expected reward to absorption starting from any state. In this setting we develop sufficient conditions under which the estimate resulting from the combined approach has a mean square error that asymptotically equals zero even when the estimate formed by using only importance sampling change of measure has infinite variance. In particular, we consider the problem of estimating the small buffer overflow probability in a queuing network, where the change of measure suggested in literature is shown to have infinite variance under certain parameters and where the appropriate combination of IS and TD method can be empirically seen to have a much faster convergence rate compared to naive simulation.

Categories and Subject Descriptors: G.3 [**Probability and Statistics**]: *Probabilistic algorithms (including Monte Carlo)*; I.6.0 [**Simulation and Modeling**]: General; I.6.1 [**Simulation and Modeling**]: Simulation Theory

General Terms: Algorithms, Performance, Theory

Additional Key Words and Phrases: Importance sampling, temporal difference methods, rare events, stochastic approximation, Markov chains, variance reduction

1. INTRODUCTION

The importance sampling (IS) variance reduction simulation techniques have been successfully used to estimate extremely small probabilities for certain

Authors' addresses: R.S. Randhawa, Graduate School of Business, Stanford University, Stanford, CA 94305; email: rsr@stanford.edu; S. Juneja, School of Technology and Computer Science, Tata Institute of Fundamental Research, Colaba, Mumbai, India—400005; email: juneja@tifr.res.in.

stochastic systems used to model communication networks, insurance processes, financial systems and reliability systems (see, e.g., Heidelberger [1995] for a survey, and Glynn and Iglehart [1989] and Juneja [2003] for an introduction). These techniques involve simulating the system under a new probability measure and then appropriately unbiasing the resultant output by weighing it with the likelihood ratio (ratio of the original and the new probability of the generated sample path). Under the importance sampling techniques, the selection of a good probability measure is critical as a wrong selection may lead to large, even infinite, variance (see, e.g., Glasserman and Wang [1997], Juneja and Shahabuddin [2001], Andradottir et al. [1995]). This drawback of the IS techniques has restricted their application to systems with a special structure such as highly reliable systems and simple queuing systems where the efficacy of the proposed IS measure is explicitly proved.

Significant literature has appeared in the last few years analyzing stochastic approximation-based simulation techniques such as the temporal difference (TD) methods to approximately solve for performance measures associated with Markov chains and more generally with Markov decision processes (see, e.g., Sutton [1988], Bertsekas and Tsitsiklis [1996], Sutton and Barto [1998]). From a classical simulation viewpoint, in the Markov chain setting the temporal difference methods may be regarded loosely as an instance of the control variates method where multiple controls are involved and the weights assigned to the controls are declining geometrically. The key difference is that, unlike in the classical settings, here the controls have mean zero only asymptotically (see Section 2 for further discussion on this).

In this paper we show that by suitably combining the TD methods with the IS techniques we get algorithms that converge to the correct solution almost surely and have asymptotically zero mean square error under a large class of importance sampling distributions, including those that may give infinite variance when applied alone. As pointed out in Glasserman and Wang [1997], Juneja and Shahabuddin [2001], and Andradottir et al. [1995], an importance sampling distribution leads to large variance in Markov chain settings as the associated likelihood ratio may increase geometrically as a function of the length of the generated path for some set of sample paths. A key feature of our approach is that we use the TD method to assign geometrically declining weights as a function of the length of the generated path to overwhelm a geometrically increasing likelihood ratio due to importance sampling.

We focus on a stochastic system modeled as a finite state space discrete time Markov chain (DTMC) with one-step transition rewards and an absorbing set of states. Our interest is in estimating the expected cumulative reward to absorption starting from any state. Note that by exploiting the regenerative structure of the Markov chain the problem of estimating steady-state measures can also be reduced to that of estimating cumulative reward till regeneration starting from the regenerative state (see, e.g., Crane and Iglehart [1975]).

In the stable Jackson network setting where each queue has a single server, the problem of efficiently simulating the probability that, starting from an empty network, the total queue length exceeds a fixed large amount before the network re-empties is an important problem that has been extensively studied.

For conciseness we refer to this probability as $P_{Overflow}$. Parekh and Walrand [1989] developed a heuristic for identifying the IS change of measure (by solving a nonlinear program) for estimating this probability in general queuing networks based on insights from large-deviations theory (see, e.g., Dembo and Zeitouni [1992] for an introduction to large-deviations theory). For notational brevity it will be useful to refer to this heuristic change of measure as $P\&W$ change of measure. Frater et al. [1991] explicitly determined this change of measure in the Jackson network setting and justified it by using time reversibility arguments. Glasserman and Kou [1995] studied the performance of the $P\&W$ change of measure in estimating $P_{Overflow}$ for two Markovian queues in tandem. They identified a range of parameters for which this change of measure works well and also where it fails (i.e., gives large variance). In this paper we also note that in a similar two-queue example, if feedback is allowed, then for certain parameter values this heuristic leads to an estimator with infinite variance.

Specifically, the contributions of this paper include the following:

(1) In a finite state space Markov chain setting, we develop sufficient conditions under which the importance sampling distribution when applied alone to estimate the expected total reward till absorption leads to infinite variance. To demonstrate its use, we consider the problem of estimating $P_{Overflow}$ for a two-queue system with feedback and show that under certain parameters the $P\&W$ change of measure satisfies these sufficient conditions and hence has infinite variance.

(2) We then show how the temporal difference methods may be combined with importance sampling (the combination is referred to as the *TDIS method*; we also propose an improvement over straightforward combination referred to as the *Improved TDIS* or the *ITDIS method*) and develop sufficient conditions under which, from each state, the estimator of the expected cumulative reward till absorption using the combined approach converges with probability 1. In addition, we show that asymptotically (as computational effort increases to infinity) the estimator has zero mean square error. In particular, it follows from our analysis that a suitable combined method may be devised to estimate $P_{Overflow}$ that uses the $P\&W$ change of measure and has asymptotically zero mean square error under every set of parameters under which the Jackson queuing network is stable.

(3) Empirically, through a small two-queue network example, we show that the combined technique converges to the correct estimate at a much faster rate compared to the naive simulation as well as the temporal difference methods. We further show that the combined technique is relatively distribution insensitive, that is, even if the change of measure is somewhat different from the $P\&W$ change of measure, convergence at a fast rate is observed as long as the event of interest is no longer rare under the new change of measure. This is a promising observation as it suggests that the combined approach may have greater applicability compared to importance sampling. In particular, the sophisticated analysis that is typically needed to come up with a good change of measure under importance sampling may not be needed as the combined approach may work well even with any

change of measure that sufficiently emphasizes (assigns high probability) the most likely paths to the rare event.

We mentioned earlier that the temporal difference methods have been used in the existing literature to solve Markov decision processes (MDP). Often, this involves fixing a policy in an MDP and estimating the cumulative reward till absorption from every initial state in the resulting Markov chain using the temporal difference methods. This suggests that the ITDIS methods developed in this paper for Markov chains are potentially useful in solving MDPs, particularly if the performance measure of interest involves rare events.

Glynn and Iglehart [1989] suggested that the stochastic approximation methods may be used to converge to the optimal importance sampling change of measure for estimating probabilities of rare events. This idea has been further analyzed in a number of papers (see, e.g., Rubinstein [1997, 1999]). Our approach differs in that we consider a fixed change of measure (for example, based on some heuristic such as the $P\&W$ change of measure) and show that, when this measure is used in conjunction with a suitable stochastic approximation based simulation technique, convergence is assured.

In a recent paper, Precup et al. [2000] also discussed combining the TD methods with the importance sampling techniques. They proposed a number of algorithms and proved that the proposed estimators are consistent. However, unlike in this paper they did not focus on convergence analysis of the variance and the mean square error to aid in identifying good and bad combinations of the importance sampling and the temporal difference methods.

In Section 2 we develop the finite state Markov chain framework and briefly review the temporal difference method in this setting, emphasizing its relation to the control variate techniques. The importance sampling techniques are reviewed in Section 3. In this section we also develop sufficient conditions under which the IS change of measure results in infinite variance. Furthermore, in Section 3 we also discuss an example of a small queuing network with feedback and show for this example that the $P\&W$ change of measure results in infinite variance under certain parameters. In Section 4 we discuss how the temporal difference methods may be combined with importance sampling and state our main result. In Section 5 we discuss the results of the simulation experiments conducted. Finally, in Section 6 we discuss potential drawbacks of this work and a possible opportunity to address these.

We prove our main result by viewing the proposed algorithms in the stochastic approximation (SA) framework and then applying the standard convergence results from the SA theory. To maintain the flow of the exposition, a brief review of the relevant SA convergence results and the proof of our main result (which uses these convergence results) are given in the Appendix.

2. MATHEMATICAL FRAMEWORK AND TEMPORAL DIFFERENCE CONTROL VARIATES

Let $(X_n : n \geq 0)$ be a DTMC where each X_n takes values in a finite state space $\mathcal{S}$. Let $P = (p_{xy} : x, y \in \mathcal{S})$ denote the transition matrix of the Markov chain. Let $G = (g(x, y) : x, y \in \mathcal{S})$ denote the one-step rewards associated with each

transition. Let $\mathcal{A} \subset \mathcal{S}$ denote the set of *absorbing* states. Let $\mathcal{I} = \mathcal{S} - \mathcal{A}$ denote the set of *interior* states. Thus, $p_{xy} = 0$ for $x \in \mathcal{A}$ and $y \in \mathcal{I}$, and also $g(x, y) = 0$ for $x \in \mathcal{A}$. We make the following mild assumption:

ASSUMPTION 2.1. *There exists a distribution μ with support $\mathcal{I}$ such that the Markov chain with transitions governed by P for states in $\mathcal{I}$ and governed by μ for every state in $\mathcal{A}$ is irreducible.*

This assumption ensures that the set $\mathcal{A}$ is reachable from any state in $\mathcal{I}$, that is, for every state $x \in \mathcal{I}$, there exists a state $y \in \mathcal{A}$ such that there exists a path of positive probability connecting x to y. Let $\tau = \inf\{n : X_n \in \mathcal{A}\}$. Note that τ is a stopping time, that is, $\{\tau = n\}$ is completely determined by observing $(X_0, X_1, \ldots, X_n)$. Let $\mathbf{P}$ denote the probability measure induced by the transition matrix P and let E denote the expectation operator corresponding to $\mathbf{P}$. We consider the problem of estimating the "value function" $J = (J(x) : x \in \mathcal{I})$, where each $J(x)$ denotes the expected cumulative reward gained by the Markov chain till absorption starting from state x. Specifically,

$$J(x) = E_x \left[\sum_{n=0}^{\tau-1} g(X_n, X_{n+1}) \right], \tag{1}$$

where the subscript x denotes that $X_0 = x$. We set $J(x) = 0$ for $x \in \mathcal{A}$.

2.1 Temporal Difference Control Variates

Under the Monte Carlo method each $J(x)$ is estimated by taking an average of independent identically distributed (i.i.d.) samples of $\sum_{n=0}^{\tau-1} g(X_n, X_{n+1})$ with $X_0 = x$ (we refer the reader to Bertsekas and Tsitsiklis [1996] for a discussion on how a single simulated sample path may be used to generate such samples for all the states visited in that path, and some nuances that need to be considered). Recall that for $m \geq \tau$, $\sum_{n=\tau}^{m} g(X_n, X_{n+1}) + J(X_{m+1}) = 0$. Thus, for each m

$$\sum_{n=0}^{m-1} g(X_n, X_{n+1}) + J(X_m) \tag{2}$$

is an unbiased estimator of $J(x)$ (for $m \geq \tau$, it equals the Monte-Carlo estimate). In fact, phrase (2) is the m-step conditional expectation of $\sum_{n=0}^{\tau-1} g(X_n, X_{n+1})$ conditioned on $(X_0, X_1, \ldots, X_m)$. The temporal difference methods consider a convex combination of these estimates where the weights assigned to each combination are declining geometrically at rate λ $(0 \leq \lambda \leq 1)$. Specifically, consider the random variable

$$(1 - \lambda) \sum_{m=1}^{\infty} \lambda^{m-1} \left[\sum_{n=0}^{m-1} g(X_n, X_{n+1}) + J(X_m) \right].$$

This is easily seen to be an unbiased estimator of $J(x)$. Changing the order of summation, this can be seen to equal

$$\sum_{n=0}^{\infty} \lambda^n g(X_n, X_{n+1}) + (1 - \lambda) \sum_{m=1}^{\infty} \lambda^{m-1} J(X_m), \tag{3}$$

which after simple manipulations can be seen to equal

$$\sum_{m=0}^{\infty} \lambda^m D_m + J(x),$$

where D_m denotes $g(X_m, X_{m+1}) + J(X_{m+1}) - J(X_m)$. It follows that

$$E_x\left(\sum_{m=0}^{\infty} \lambda^m D_m\right) = 0. \tag{4}$$

It is also easily seen that $E_x(D_m) = 0$ for $x \in I$. To see this, note that

$$J(X_m) = E_x[g(X_m, X_{m+1}) + J(X_{m+1})|(X_1, \ldots, X_m), m < \tau] \text{ almost surely}$$

and hence $E_x[D_m|m < \tau] = 0$ almost surely. Also note that $E_x[D_m|m \geq \tau] = 0$ almost surely trivially as $J(x) = g(x, y) = 0$ for all $x \in \mathcal{A}$. Each D_m may be defined as a step m temporal difference control variate.

The *TD* methods exploit the relationship in Equation (4). Note that generating D_m via simulation requires the knowledge of $J(X_m)$ and $J(X_{m+1})$, which are typically a priori not known. The *TD* methods use the stochastic approximation framework to update the estimates of $J(x)$ for each $x \in \mathcal{I}$ along a generated sample path. These estimates are then used to generate estimates of D_m. The *TD* methods are a family of algorithms parameterized by $0 \leq \lambda \leq 1$. For a given λ, the *TD*(λ) algorithm is as follows (below, let $\mathcal{F}_t$ denote the information corresponding to all the random variables generated before iteration t is initiated):

(1) Select initial estimates $(J_0(x) : x \in \mathcal{I})$ arbitrarily (e.g., $J_0(x) = 0$ for $x \in \mathcal{I}$). Set $t = 0$.

(2) At any iteration t, select $X_{0,t}$ using a distribution μ guaranteed by Assumption 2.1 and using simulation under **P** generate random variables $(X_{1,t}, \ldots, X_{\tau_t,t})$ of the Markov chain till absorption, where τ_t denotes the first time the absorbing set $\mathcal{A}$ is hit (through simulation one generates realizations of collections of random variables (rv), not random variables; however, we abuse the convention and use rv to denote the resulting realizations in the algorithm to avoid unnecessary extra notation).

(3) For each state x visited for the first time at the transition n of the generated random variables (that is, $X_{n,t} = x$, and $X_{j,t} \neq x$ for $j < n$), the value function is updated as follows:

$$J_{t+1}(x) = J_t(x) + \gamma_t(x) \sum_{m=n}^{\tau_t - 1} \lambda^{m-n} D_{m,t}, \tag{5}$$

where $(\gamma_t(\cdot) : t \geq 1)$ are the nonnegative step-size parameters satisfying the usual conditions of decreasing step-sizes in stochastic approximation literature, that is, $\sum_{t=0}^{\infty} \gamma_t(x) = \infty$ and $\sum_{t=0}^{\infty} \gamma_t^2(x) < \infty$ almost surely for each x (they may depend upon $\mathcal{F}_t$), and $D_{m,t}$ equals $g(X_{m,t}, X_{m+1,t}) + J_t(X_{m+1,t}) - J_t(X_{m,t})$ (it is a sample of appropriate D_m and may be biased as the vector J_t may be a biased estimator of vector J). Note that if a state is visited multiple times, this update is performed only once corresponding to the

first visit. For all states x not visited by the generated sequence of random variables set $J_{t+1}(x) = J_t(x)$.

(4) Set $t = t + 1$ and repeat step 2.

Let $t(x)$ denote the number of visits to state x up till iteration t. Suppose that $\gamma_t(x) = 1/t(x)$, $J_0(x) = 0$, and $\lambda = 1$. Further, suppose that x is visited at transition n during iteration t for the first time in this iteration. Then, $J_{t+1}(x)$ equals the average of $t(x)$ i.i.d. samples of total cumulative rewards till absorption starting from state x, that is, the TD method is similar to the usual Monte Carlo method of estimating $J(x)$. To see this, note that in this case, given that $X_{n,t} = x$,

$$\sum_{m=n}^{\tau_t-1} \lambda^{m-n} D_{m,t} = \sum_{m=n}^{\tau_t-1} g(X_{m,t}, X_{m+1,t}) - J_t(X_{n,t}),$$

and thus

$$J_{t+1}(x) = \frac{J_t(x)(t(x)-1) + \sum_{m=n}^{\tau_t-1} g(X_{m,t}, X_{m+1,t})}{t(x)}.$$

Through an inductive argument it follows that $J_{t+1}(x)$ is an average of $t(x)$ i.i.d. samples of the total cumulative rewards.

The update procedure shown in Equation (5) is referred to as the *off-line update procedure* as the complete sample path at time t is generated and then the update is performed for all the states visited along the path. In this paper, we focus on these off-line methods due to the relative simplicity of analysis. We refer the reader to Bertsekas and Tsitsiklis [1996] for a discussion on very similar and more efficient on-line updating methods and also for the proof of convergence of the above algorithm.

Typically, under the $TD(\lambda)$ method, empirically it is observed that there exists an optimal λ^* between 0 and 1 such that the $TD(\lambda^*)$ method performs better than the $TD(\lambda)$ methods for $\lambda \neq \lambda^*$. In particular, the greater the difference $|\lambda - \lambda^*|$ the greater the performance improvement of the $TD(\lambda^*)$ over the $TD(\lambda)$ methods (see, e.g., Singh and Dayan [1998]). Thus, it may perform significantly better than the Monte Carlo or the suitable TD(1) method.

Remark 2.2. In Heidelberger [1980a, 1980b, 1977] and Andradottir et al. [1993], the problem of estimating the total cumulative reward in a regenerative cycle is considered in an irreducible finite state Markov chain setting (this is useful in estimating the steady state average reward via the regenerative ratio representation). They used different types of m step conditional expectations as control variates to help speed up the simulation. Specifically, suppose again that $g(\cdot, \cdot)$ denotes one-step rewards, x is the regenerative state, and $\tilde{\tau}$ denotes the time to return to state x. In this setting they considered the problem of estimating $E_x(\sum_{n=0}^{\tilde{\tau}-1} g(X_n, X_{n+1}))$. The essential idea is that in many cases it may be feasible to cheaply numerically compute m-step conditional expectation $E(g(X_m, X_{m+1})|X_0 = x)$ for all x. It can be seen from Heidelberger [1980a] that

$$E_x\left(\sum_{n=0}^{\tilde{\tau}-1} g(X_n, X_{n+1})\right) = E_x\left(\sum_{n=0}^{\tilde{\tau}-1} E(g(X_{n+m}, X_{n+m+1})|X_n)\right).$$

Thus, for each m, $\sum_{n=0}^{\tilde{\tau}-1} E(g(X_{n+m}, X_{n+m+1})|X_n)$ provides an alternative estimator of total reward in a regenerative cycle. Heidelberger [1980a, 1980b, 1977] and Andradottir et al. [1993] also discussed how to determine the minimum variance linear combination of these estimates and related techniques.

Note that while these techniques are different from the TD methods, it may indeed be feasible to combine them with the TD methods (and with importance sampling discussed later) by replacing each $g(X_m, X_{m+1})$ in the temporal difference control variate D_m by a suitable weighted average of $g(X_m, X_{m+1})$ and $(E(g(X_m, X_{m+1})|X_n) : n \leq m)$. However, to maintain focus we do not explore this idea further in this paper.

3. IMPORTANCE SAMPLING

Under the importance sampling techniques the Markov chain is simulated using a different probability distribution with transition matrix $P' = (p'_{xy} : x, y \in \mathcal{S})$, which has the property that P is absolutely continuous with respect to P', that is, for all $x, y \in \mathcal{S}$, $p'_{xy} = 0$ implies $p_{xy} = 0$. Let $\mathbf{P}'$ denote the distribution induced by P' and let $\bar{E}$ denote the corresponding expectation operator. Then $J(x)$ may be reexpressed as

$$J(x) = \bar{E}_x \left[\left(\sum_{n=0}^{\tau-1} g(X_n, X_{n+1}) \right) L(X_0, X_1, \ldots, X_\tau) \right], \tag{6}$$

where

$$L(X_0, X_1, \ldots, X_\tau) = \prod_{n=0}^{\tau-1} \frac{p_{X_n, X_{n+1}}}{p'_{X_n, X_{n+1}}}.$$

See Glynn and Iglehart [1989] for a detailed description. For brevity denote rv $L(X_0, X_1, \ldots, X_\tau)$ by L. In the ensuing discussion fix $X_0 = x$ and suppress x from the notation. Let K be a rv denoting the quantity $\sum_{n=0}^{\tau-1} g(X_n, X_{n+1})$. As mentioned earlier, $J = E[K]$.

Under importance sampling, the probability measure $\mathbf{P}'$ is used to generate i.i.d. samples $(K_1, L_1), (K_2, L_2), \ldots, (K_n, L_n)$ of (K, L) and the new estimator is $\hat{J}_{P'} = \frac{1}{n} \sum_{j=1}^{n} K_j * L_j$. The variance $\sigma^2(P')$ of the rv $K * L$ is given by

$$\sigma^2(P') = \bar{E}[K^2 * L^2] - (\bar{E}[K * L])^2 = \bar{E}[K^2 * L^2] - J^2. \tag{7}$$

As the value of J is a constant, the variance is characterized by the second moment $\bar{E}[K^2 * L^2]$. Some notation is needed to examine the second moment term more closely. Let ω denote a generic realization of collection of rv $(X_1, X_2, \ldots, X_\tau)$. Let Ω denote the set of all such realizations. Let $P(\omega)$ and $P'(\omega)$ denote the probability of occurrence of ω under $\mathbf{P}$ and $\mathbf{P}'$, respectively. Let $K(\omega)$ denote the value of rv K along ω. Then,

$$\bar{E}[K^2 * L^2] = \sum_{\omega \in \Omega} \frac{K^2(\omega) P^2(\omega)}{P'^2(\omega)} P'(\omega) = \sum_{\omega \in \Omega} \frac{K^2(\omega) P^2(\omega)}{P'(\omega)}. \tag{8}$$

The aim of importance sampling is to find an implementable $\mathbf{P}'$ so that this term is minimized or at the very least it is less than the second moment

corresponding to naive simulation (i.e., $\sum_{\omega\in\Omega} K^2(\omega)P(\omega)$). However, as mentioned in the Introduction, this is often not an easy task and carelessly chosen $\mathbf{P}'$ may in fact *increase* the variance.

To further illustrate this, in Assumption 3.1 in Section 3.1 we state sufficient conditions under which the importance sampling estimator has infinite second moment and hence infinite variance. In Section 3.2, these conditions are used to show examples where the $P\&W$ measure leads to infinite variance in simple Jackson network settings.

3.1 Sufficient Conditions for Infinite Variance

Let $I(\tau \le n)$ be an indicator function (it takes value 1 if the path to the absorbing set is of length $\le n$ and takes value 0 otherwise). Note that $\bar{E}(K^2L^2I(\tau \le n))$ is a bounded quantity. To see this, let $\alpha = \max_{x,y} \frac{p_{xy}}{p'_{xy}}$. This is finite due to the finite state space of the Markov chain and since P is absolutely continuous with respect to P'. Similarly, $c = \max_{x,y} g(x, y)$ is finite. Then, since $K\,I(\tau \le n) \le nc$ and $L\,I(\tau \le n) \le \alpha^n$ (note that $\alpha \ge 1$) we have

$$\bar{E}(K^2L^2I(\tau \le n)) \le (nc)^2\alpha^{2n}.$$

Thus, infinite second moment may occur only if paths of unbounded length are considered. In particular, in the finite state space setting it may result from paths that have "cycles."

A cycle is an ordered sequence of two or more states where the first and the last state are identical (see [Juneja 2001] for application of the "cyclic approach" to determine good IS distributions in some rare event settings). Thus, a cycle may contain many cycles within it. A path connecting two states is called a "direct" path if it does not include cycles. Consider a cycle $\mathcal{C} = (x_0, x_1, x_2, \ldots, x_{n-1}, x_n)$ (i.e., $x_0 = x_n$). Let $P(\mathcal{C})$ denote its probability under the original probability distribution $\mathbf{P}$, that is, $P(\mathcal{C}) = \prod_{j=0}^{n-1} p_{x_j x_{j+1}}$, and $P'(\mathcal{C})$ denote the probability of cycle $\mathcal{C}$ under the new probability distribution $\mathbf{P}'$, that is, $P'(\mathcal{C}) = \prod_{j=0}^{n-1} p'_{x_j x_{j+1}}$. Similarly, for any direct path $\mathcal{D}$ let $P(\mathcal{D})$ and $P'(\mathcal{D})$ denote the probability of the direct path under $\mathbf{P}$ and $\mathbf{P}'$, respectively.

Any realization $(x_0, x_1, \ldots, x_n)$ is referred to as *probable* if $\prod_{j=0}^{n-1} p_{x_j x_{j+1}} > 0$. A path or a cycle is said to *belong to a set* A if all states comprising them belong to A. For any state $y \in \mathcal{I}$, let C_y denote a collection of probable cycles in $\mathcal{I}$ whose first and last states are y but which do not include y in the remaining states. Note that there may be infinite number of cycles in C_y. To see this, suppose that (y, a, y) and (a, b, a) are probable cycles for $a, b \in \mathcal{I}$. Then, the cycles (y, a, y), (y, a, b, a, y), (y, a, b, a, b, a, y), and so on all belong to C_y.

We make the following assumption regarding the Markov chain under consideration:

ASSUMPTION 3.1. *The following conditions hold:*

(1) *There exists a set* $\mathcal{W} \subset \mathcal{A}$ *such that* $\mathbf{P}(X_\tau \in \mathcal{W}) > 0$ *and* $K \ge \delta$ *for a constant* $\delta > 0$ *on the set* $\{X_\tau \in \mathcal{W}\}$.

(2) *There exists a state* $y \in \mathcal{I}$ *and for some* m *there exist cycles* $(\mathcal{C}_j : j \le m)$ *in*

C_y such that

$$\sum_{j \le m} \frac{P(\mathcal{C}_j)^2}{P'(\mathcal{C}_j)} \ge 1.$$

(3) *There exists a probable direct path $\mathcal{D}_1$ in $\mathcal{I}$ connecting the initial state x to y and a probable direct path $\mathcal{D}_2$ in $\mathcal{I}$ connecting state y to a state in $\mathcal{W}$.*

LEMMA 3.2. *Under Assumption* (3.1), *the estimator $\hat{J}_{P'}$ has infinite variance.*

PROOF. Let $(n_j : j \le m)$ denote a vector of nonnegative integers and let $n = \sum_{j=1}^{m} n_j$. Now consider a sample path ω comprising a direct path $\mathcal{D}_1$ to y, n_j cycles $\mathcal{C}_j$ for $(j \le m)$ in any order and a direct path $\mathcal{D}_2$ from y to a state in $\mathcal{W}$. For this sample path

$$\frac{P^2(\omega)}{P'(\omega)} = \frac{P^2(\mathcal{D}_1)P^2(\mathcal{D}_2)}{P'(\mathcal{D}_1)P'(\mathcal{D}_2)} \prod_{j \le m} \left(\frac{P(\mathcal{C}_j)^2}{P'(\mathcal{C}_j)} \right)^{n_j}.$$

Note that, for each fixed n, we have $\frac{n!}{n_1!n_2!\cdots n_m!}$ such distinct paths with n_1 cycles $\mathcal{C}_1$, n_2 cycles $\mathcal{C}_2$, and so on, when all possible sequencing of the cycles is considered. In particular, considering all such paths, for all n_i such that $n_1 + n_2 + \cdots + n_m = n$ and for all $n \ge 1$, and noting that $K \ge \delta$ along such paths, we get

$$\begin{aligned}
&\sum_{\omega \in \Omega} \frac{K^2(\omega)P^2(\omega)}{P'(\omega)} \\
&\quad \ge \delta^2 \frac{P^2(\mathcal{D}_1)P^2(\mathcal{D}_2)}{P'(\mathcal{D}_1)P'(\mathcal{D}_2)} \sum_{n=1}^{\infty} \sum_{(\sum_{j=1}^{m} n_j = n)} \frac{n!}{n_1!n_2!\cdots n_m!} \prod_{j=1}^{m} \left(\frac{P(\mathcal{C}_j)^2}{P'(\mathcal{C}_j)} \right)^{n_j} \qquad (9) \\
&\quad = \delta^2 \frac{P^2(\mathcal{D}_1)P^2(\mathcal{D}_2)}{P'(\mathcal{D}_1)P'(\mathcal{D}_2)} \sum_{n=1}^{\infty} \left(\sum_{j=1}^{m} \frac{P(\mathcal{C}_j)^2}{P'(\mathcal{C}_j)} \right)^{n}, \qquad (10)
\end{aligned}$$

where Equation (10) follows as the last summation in Equation (9) is simply the multinomial expansion of $(\sum_{j=1}^{m} \frac{P(\mathcal{C}_j)^2}{P'(\mathcal{C}_j)})^n$. Therefore, if $\sum_{j=1}^{m} \frac{P(\mathcal{C}_j)^2}{P(\mathcal{C}_j)} \ge 1$, the importance sampling estimator has infinite second moment and infinite variance. □

3.2 Examples of Infinite Variance in a Jackson Network

We now use Lemma 3.2 to show that the *P&W* change of measure may give infinite variance even in a simple two-queue Jackson network setting where feedback is allowed. Let (ν_1, ν_2) denote the arrival rates to queues 1 and 2, respectively. Let (μ_1, μ_2) denote the service rates at queues 1 and 2, respectively. Let $(b_{ij} : i = 1, 2; j = 1, 2, e)$ denote the transition probabilities associated with this queuing system, where b_{ie} denotes the probability that a customer leaving queue i exits the system. Also, for $i = 1, 2$, let γ_i denote the total arrival rate to queue i.

Note that if Q_{in} denotes the length of queue i at instant n of system state change (due to an arrival or departure), then $((Q_{1n}, Q_{2n}) : n \ge 0)$ is a DTMC

taking values in the nonnegative quadrant. Let $\mathcal{O}$ denote a set containing a single state $(0, 0)$ and let $\mathcal{R} = ((x_1, x_2) : x_1 + x_2 = B; x_1, x_2 \in Z^+)$, where Z^+ is a set of nonnegative integers. Thus, in a stable queuing system (i.e., $\gamma_i < \mu_i$, $i = 1, 2$) $\mathcal{O}$ denotes an attractor set (set of state(s) visited frequently by the Markov chain) while, for large B, $\mathcal{R}$ denotes a rare set. In this setting, efficiently estimating the probability that, once the system becomes busy, it hits $\mathcal{R}$ before it hits $\mathcal{O}$ is important as its efficient estimation is key to the efficient estimation of the practically important steady state loss probability of an associated queuing system (see, e.g., Parekh and Walrand [1989], Heidelberger [1995]) with a common buffer of size B.

The estimation of this probability is incorporated in our framework as follows: the state space of the Markov chain $\mathcal{S} = ((x_1, x_2) : x_1 + x_2 \leq B; x_1, x_2 \in Z^+)$. The absorbing set $\mathcal{A}$ equals $\mathcal{O} \cup \mathcal{R}$. Interior set $\mathcal{I}$ equals $\mathcal{S} - \mathcal{A}$. The one-step rewards are $g(u, v) = 1$ if $u \in \mathcal{I}$, $v \in \mathcal{R}$, and $g(u, v) = 0$ otherwise. The starting state is $(1, 0)$ with probability $\nu_1/(\nu_1 + \nu_2)$ or $(0, 1)$ with probability $\nu_2/(\nu_1 + \nu_2)$. To see that condition 1 of Assumption 3.1 is valid in this setup, note that all the paths starting from either $(1, 0)$ or $(0, 1)$ that hit $\mathcal{R}$ before they hit $\mathcal{O}$ have a total reward equal to 1. Also note that condition 3 of Assumption 3.1 holds trivially. In Examples 3.3 and 3.4 we list some parameters under which the $P\&W$ change of measure gives infinite variance. In Example 3.3, condition 2 of Assumption 3.1 is satisfied with $m = 1$, while in Example 3.4 it holds with $m = 2$. The equations used to compute this change of measure are taken from Frater et al. [1991] and are listed in Section A.4 of the Appendix.

Example 3.3. Suppose that $(\nu_1, \nu_2, \mu_1, \mu_2) = (0.09, 0.01, 0.77, 0.13)$. The transition probabilities are $(b_{11}, b_{12}, b_{1e}) = (0, 0.05, 0.95)$ and $(b_{21}, b_{22}, b_{2e}) = (0.9, 0, 0.1)$. The traffic intensities are $\rho_1 = 0.135$ and $\rho_2 = 0.117$ for the two queues, respectively. Under the $P\&W$ change of measure, $(\nu_1', \nu_2', \mu_1', \mu_2') = (0.67, 0.07, 0.13, 0.13)$. The new transition probabilities are

$$(b_{11}', b_{12}', b_{1e}') = (0, 0.26, 0.74),$$

and $(b_{21}', b_{22}', b_{2e}') = (0.985, 0, 0.015)$. The new traffic intensities are $\rho_1' = 5.76$ and $\rho_2' = 0.79$. For this system consider a probable cycle $\mathcal{C} = ((0, 1), (1, 1), (0, 1))$. For this cycle $P(\mathcal{C}) = 0.29$ and $P'(\mathcal{C}) = 0.079$. Thus, $\frac{P(\mathcal{C})^2}{P'(\mathcal{C})} = 1.07$. Therefore, Assumption 3.1 holds and from Lemma 3.2 it follows that the IS estimator has infinite variance.

Example 3.4. Suppose that $(\nu_1, \nu_2, \mu_1, \mu_2) = (0.06, 0.01, 0.79, 0.14)$ and let the transition probabilities $(b_{11}, b_{12}, b_{1e}) = (0, 0.05, 0.95)$ and $(b_{21}, b_{22}, b_{2e}) = (0.9, 0, 0.1)$. The traffic intensities are $\rho_1 = 0.0976$ and $\rho_2 = 0.0967$. Under the $P\&W$ change of measure, $(\nu_1', \nu_2', \mu_1', \mu_2') = (0.66, 0.1, 0.1, 0.14)$. The new transition probabilities are $(b_{11}', b_{12}', b_{1e}') = (0, 0.33, 0.67)$ and $(b_{21}', b_{22}', b_{2e}') = (0.99, 0, 0.01)$. The new traffic intensities are $\rho_1' = 7.23$ and $\rho_2' = 0.9$. Now consider the cycles $\mathcal{C}_1 = ((0, 1), (1, 1), (0, 1))$ and $\mathcal{C}_2 = ((0, 1), (1, 0), (0, 1))$. For these cycles, we see that $P(\mathcal{C}_1) = 0.22$, $P(\mathcal{C}_2) = 0.027$, $P'(\mathcal{C}_1) = 0.053$ and $P'(\mathcal{C}_1) = 0.006$. Observe that $\frac{P(\mathcal{C}_1)^2}{P'(\mathcal{C}_1)} + \frac{P(\mathcal{C}_2)^2}{P'(\mathcal{C}_2)} = 1.03$. Therefore, from Lemma 3.2 we again conclude that the resultant estimator has infinite variance.

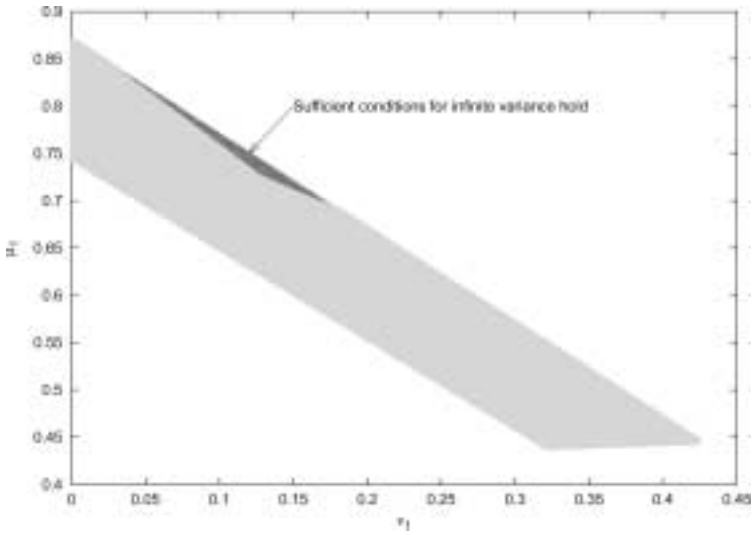

Fig. 1. The darkened region shows some parameters for which Assumption 3.1 holds under the $P\&W$ change of measure and hence the IS estimator has infinite variance. Note that there may be other parameters on the graph where the variance is infinite. The entire shaded region denotes the parameters for which the network is stable.

In Figure 1 we plot the range of parameters where, for the cycles described in Example 3.4, $\frac{P(\mathcal{C}_1)^2}{P'(\mathcal{C}_1)} + \frac{P(\mathcal{C}_2)^2}{P'(\mathcal{C}_2)} \geq 1$. We fix the values of the transition probabilities and the service rate of queue 2 and then vary the values of ν_1, ν_2, and μ_1 in such a way that their sum remains constant and the resulting network is still stable. The darkened region in the figure indicates the values of μ_1 and ν_1 for which the importance sampling estimator has infinite variance, while the entire shaded region displayed in the figure represents the set of values of μ_1 and ν_1 for which the network is stable.

4. THE COMBINED TDIS METHOD

In Section 4.1, we develop the TDIS method for estimating expected total reward from each state till absorption by combining the temporal difference methods with importance sampling. A potential drawback of this method may be that, even in the scenario where all the one-step rewards are nonnegative, the combined method may lead to negative estimates of expected total reward till absorption (this is demonstrated later in the section). To remedy this, in Section 4.2 we propose an Improved TDIS method (referred to as ITDIS) so that the estimators of the expectations are always nonnegative. Finally, in Section 4.3, we state our main theorem that gives sufficient conditions under which the proposed methods converge to the correct values.

4.1 The TDIS Method

The TDIS method relies on the following lemma:

LEMMA 4.1. *For each* $x \in \mathcal{I}$,

$$\bar{E}_x \left[\sum_{m=0}^{\tau-1} \lambda^m D_m L(X_0, \ldots, X_{m+1}) \right] = 0.$$

PROOF. First note that $\bar{E}_x(D_m L(X_0, \ldots, X_{m+1})) = 0$. Thus, we simply need to justify the interchange of the summation and the expectation term in

$$\bar{E}_x \left[\sum_{m=0}^{\infty} \lambda^m D_m L(X_0, \ldots, X_{m+1}) I(m < \tau) \right]. \tag{11}$$

Note that the rv under the expectation operator in phrase (11) is upper bounded by

$$\sum_{m=0}^{\infty} \lambda^m |D_m| L(X_0, \ldots, X_{m+1}) I(m < \tau).$$

Thus, the interchange of the summation and the expectation in phrase (11) is justified if

$$\bar{E}_x \left[\sum_{m=0}^{\infty} \lambda^m |D_m| L(X_0, \ldots, X_{m+1}) I(m < \tau) \right] < \infty. \tag{12}$$

This follows since, due to the nonnegativity of the terms involved, the left-hand side in Equation (12) equals:

$$\sum_{m=0}^{\infty} \bar{E}_x [\lambda^m |D_m| L(X_0, \ldots, X_{m+1}) I(m < \tau)] = \sum_{m=0}^{\infty} E_x [\lambda^m |D_m| I(m < \tau)] \tag{13}$$

(due to Fubini's theorem; see Billingsley [1995]).

This is finite for $\lambda < 1$ as each $|D_m|$ is uniformly bounded by a constant (due to the finite state space). For $\lambda = 1$, this follows by also noting that $E_x \tau < \infty$. □

Lemma 4.1 suggests the following TDIS algorithm for estimating $(J(x) : x \in \mathcal{I})$ (again let $\mathcal{F}_t$ denote the information corresponding to all the rv generated before initiation of iteration t):

(1) Select initial estimates $(J_0(x) : x \in \mathcal{I})$ arbitrarily. Set $t = 0$.

(2) At any iteration t, select $X_{0,t}$ using any distribution μ guaranteed by Assumption 2.1 and using simulation under $\mathbf{P}'$, generate rv $(X_{1,t}, \ldots, X_{\tau_t,t})$ of the Markov chain till absorption.

(3) For each state x visited for the first time at transition n of the generated rv, update the value function as follows:

$$J_{t+1}(x) = J_t(x) + \gamma_t(x) \sum_{m=n}^{\tau_t - 1} \lambda^{m-n} D_{m,t} \mathcal{L}_{n,m,t}, \tag{14}$$

where $D_{m,t}$ has the same interpretation as in the TD algorithm and $\mathcal{L}_{n,m,t}$ equals

$$L(X_{n,t}, \ldots, X_{m+1,t}).$$

For all states x not visited in iteration t, set $J_{t+1}(x) = J_t(x)$.

(4) Set $t = t + 1$ and repeat step 2.

The convergence result for this algorithm is discussed later in Section 4.3. However, a potential drawback of this approach is that it may give negative estimates even if all the transition rewards are nonnegative. To see this clearly consider the following simple scenario: let $x \in \mathcal{I}$ and $y \in \mathcal{A}$ be two states such that $p_{xy} > 0$. Let $J_t(x)$ denote the estimate of the value function at state x after iteration $t-1$. Suppose that it is nonnegative. Also suppose that $g(x, y) = 0$. Now consider the case where a sample path generated at iteration t ends at y through x, that is, the last transition on the sample path is from x to y. Further suppose that the generated path visited x only once. Let $L_{xy} = p_{xy}/p'_{xy}$. Hence, the update at state x is

$$\begin{aligned} J_{t+1}(x) &= J_t(x) + \gamma_t(x)(g(x, y) + J_t(y) - J_t(x)) * L_{xy} \\ &= (1 - \gamma_t(x)L_{xy})\, J_t(x). \end{aligned}$$

(since $J_t(y) = 0$ for all t). Hence, if $\gamma_t(x)L_{xy} > 1$, the estimate is negative.

To circumvent this problem, we combine the TD method with importance sampling in a slightly different manner using the Improved TDIS algorithm.

4.2 Improved TDIS Algorithm

Let $U_m = g(X_m, X_{m+1}) + (1-\lambda) J(X_{m+1})$. From phrase (3) it follows that $J(x) = E_x[\sum_{m=0}^{\tau-1} \lambda^m U_m]$. Arguing as in Lemma 4.1, it can be seen that

$$J(x) = \bar{E}_x \left[\sum_{m=0}^{\tau-1} \lambda^m U_m L(X_0, \ldots, X_{m+1}) \right].$$

The ITDIS algorithm relies on this observation and differs from the TDIS algorithm only in that the update at step 3 is governed by

$$J_{t+1}(x) = (1 - \gamma_t(x))\, J_t(x) + \gamma_t(x) \left(\sum_{m=n}^{\tau_t - 1} \lambda^{m-n} U_{m,t} \mathcal{L}_{n,m,t} \right), \tag{15}$$

where $U_{m,t}$ equals $g(X_{m,t}, X_{m+1,t}) + (1-\lambda) J_t(X_{m+1,t})$.

Note that if one-step rewards are nonnegative, $\gamma_t(\cdot) \leq 1$ for all t, and $(J_0(x) : x \in \mathcal{I})$ are assigned nonnegative values then the estimators remain nonnegative at each iteration of the ITDIS algorithm. This suggests that the ITDIS algorithm may perform better than the TDIS algorithm in practice. Our experiments in Section 5 corroborate this suggestion.

4.3 Convergence of TDIS and ITDIS Algorithms

In the proof for convergence we restrict our analysis to all $\lambda \leq 1$ that satisfy the following relation:

$$I(m < \tau)\lambda^m * L(X_0, \ldots, X_{m+1}) \leq C \text{ almost surely} \tag{16}$$

for all m and for $X_0 \in \mathcal{I}$.

Note that for any $\mathbf{P}'$ such that $\mathbf{P}$ is absolutely continuous with respect to $\mathbf{P}'$, we can find a positive λ that satisfies relation (16). To see this, recall that $\alpha = \max_{x \in \mathcal{I}, y \in \mathcal{S}} \frac{p_{xy}}{p'_{xy}}$. Then, $I(m < \tau)L(X_0, \ldots, X_{m+1}) \leq \alpha^{m+1}$. If we choose $\lambda \leq 1/\alpha$,

relation (16) is satisfied. Note that $\lambda = 0$ always works! However, as discussed earlier, it has been empirically observed that a small value of λ typically leads to a higher bias compared to values nearer to the "best" value. Fortunately, a higher value of λ may also satisfy relation (16). To see this additional notation is needed.

For any cycle $\mathcal{C}$ such that $P'(\mathcal{C}) > 0$, let $L(\mathcal{C})$ equal the ratio $P(\mathcal{C})/P'(\mathcal{C})$. A cycle $\mathcal{C} = (x_0, \ldots, x_n)$ is referred to as a *simple* cycle if it does not include other cycles in it, that is, if $x_0, x_1, \ldots, x_{n-1}$ are all distinct states and $x_0 = x_n$. Let $n(\mathcal{C})$ denote the number of transition pairs in a cycle $\mathcal{C}$. Thus, for $\mathcal{C} = (x_0, \ldots, x_n)$, $n(\mathcal{C}) = n$. Let $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_m$ denote the total number of distinct simple cycles in $\mathcal{I}$. Due to finite state space, m is finite. Let $\beta = \max_{i \le m} L(\mathcal{C}_i)^{\frac{1}{n(\mathcal{C}_i)}}$.

LEMMA 4.2. *For*

$$\lambda \le \frac{1}{\beta},$$

relation (16) *holds.*

PROOF. Consider any probable path $(x_0, x_1, \ldots, x_m, x_{m+1})$ of the Markov chain where $x_i \in \mathcal{I}$ for $i \le m$. Clearly, due to the Markov property, we may write the probability of this path (both under $\mathbf{P}$ and $\mathbf{P}'$) as the product of the probability of the direct path and the probability of the cycles along the path. In particular, the likelihood ratio of the path may be written as the product of the likelihood ratio of the direct path and the likelihood of the cycles along the path.

Note that due to the finite state space, the number of distinct direct paths of any length in the state space is finite. Let γ denote the maximum value of the likelihood ratio along all probable direct paths. Also note that the likelihood ratio of any simple cycle of length n is upper bounded by β^n. It is easy to see that the likelihood ratio of any cycle may be written as the product of likelihood ratio of the simple cycles comprising it. Hence if the number of transitions in the cycle are n then again its likelihood ratio is upper bounded by β^n. From this it is easy to see that $L(X_0, X_1, \ldots, X_{m+1}) \le \gamma\beta^{m+1}$. The result follows by letting $C = \gamma * \beta$. □

Note that the efficient computation of β for important Markov chains such as Jackson networks is an interesting issue that needs further analysis. However, this is beyond the scope of this paper. Now we state our main result:

THEOREM 4.3. *For the Markov chain* $(X_n : n \ge 0)$, *when Assumption* (2.1) *holds and* λ *satisfies relation* (16), *the sequence* $(J_t(\cdot) : t \ge 0)$ *under both the TDIS and the ITDIS algorithms converge to* $J(\cdot)$ *with probability* 1. *In addition, the mean square error under these algorithms converges to zero.*

The proof of Theorem 4.3 is given in the Appendix. It involves representing the proposed algorithms in Robbins-Monro stochastic approximation framework and then proving that the convergence conditions are satisfied. We also extensively use the results used in the proof of off-line temporal difference method given in Bertsekas and Tsitsiklis [1996]. In the Appendix, we briefly

review the convergence conditions for the Robbins-Monro stochastic approximation scheme before proving our results. We also appropriately modify our terminology to match that used by Bertsekas and Tsitsiklis [1996] to aid in using their results.

5. SIMULATION RESULTS

In our simulations we focus on estimating $P_{Overflow}$, that is, the probability that once a Jackson network becomes busy, the total network population hits B before the network reempties. We conduct simulations on two examples, each a Jackson network with two queues and a single server at each queue. The first corresponds to that described in Example 3.3 where the $P\&W$ change of measure leads to infinite variance. The second corresponds to a tandem queue where the $P\&W$ change of measure is proved to be asymptotically optimal in Glasserman and Kou [1995]. The second model is considered simply to demonstrate that the combined method may be expected to work at least as well as the method where only importance sampling is used (referred to as the *IS method*) under all conditions. The importance sampling distribution used to simulate these networks corresponds to the $P\&W$ change of measure. For the first example we run extensive simulations using the ITDIS method to empirically determine a good set of step-sizes and the λ value. We also test the performance of the ITDIS method when the underlying change of measure is obtained by perturbing the $P\&W$ change of measure.

We compare different algorithms and each algorithm for different set of parameters by comparing their estimated MSEs resulting from a fixed computational budget. To estimate the MSE of a given method, we first numerically determine the exact value of the probability (using value iteration). Then 20 independent replications are generated from that method with each replication using one-twentieth of the fixed computational budget. The average of the square errors resulting from each replication provides the estimated MSE.

In both the examples, the initial state for each iteration in the TD, the TDIS, and the ITDIS methods was taken to be (0, 0), that is, both queues empty. Initial estimates $(J_0(x) : x \in \mathcal{I})$ were set to zero in these methods.

5.1 First Example

Consider Example 3.3 (described in Section 3.2) with $B = 8$. $P_{Overflow}$ for $B = 8$ is numerically found to equal 3.37×10^{-6}. To compare the ITDIS method to the naive simulation method (referred to as the *NS method*) and other algorithms we need to select λ and the step-sizes $(\gamma_t(\cdot) : t \geq 1)$ for its implementation.

To select a good value of λ note that, in this network, the cycle $\mathcal{C}$ equal to $((0, 1), (0, 2), (0, 1))$ has $L_{\mathcal{C}} = P_{\mathcal{C}}/P'_{\mathcal{C}} = 13.63$. Also, for this cycle $n(\mathcal{C}) = 2$. It can be shown by enumerating all simple cycles that β equals $L_{\mathcal{C}}^{\frac{1}{n(\mathcal{C})}} = 3.692$. Hence, Lemma 3 holds for $\lambda \leq \frac{1}{3.69} = 0.271$ and the MSE corresponding to the ITDIS and the TDIS methods converges to 0. Thus, to test different step-sizes $(\gamma_t(\cdot) : t \geq 1)$ we set $\lambda = 0.27$ (sensitivity to λ is studied later).

Table I. Estimated MSE for ITDIS Algorithm for Different Values of α and N at $\pi = 0.2$

		N		
		1	10	50
α	0.6	1.01×10^{-13}	1.54×10^{-14}	5.17×10^{-14}
	0.7	5.43×10^{-13}	9.13×10^{-15}	1.28×10^{-14}
	0.8	2.54×10^{-12}	1.90×10^{-14}	7.57×10^{-15}
	0.9	8.26×10^{-12}	1.13×10^{-13}	6.09×10^{-15}
	1.0	1.10×10^{-11}	5.61×10^{-13}	6.09×10^{-15}

Table II. Estimated MSE for ITDIS Algorithm for Different Values of α and N at $\pi = 0.5$

		N		
		1	10	50
α	0.6	8.71×10^{-15}	6.40×10^{-14}	6.67×10^{-14}
	0.7	3.11×10^{-14}	2.91×10^{-14}	5.29×10^{-14}
	0.8	2.84×10^{-13}	5.61×10^{-15}	3.77×10^{-14}
	0.9	1.97×10^{-12}	1.13×10^{-14}	1.13×10^{-14}
	1.0	7.06×10^{-12}	1.70×10^{-14}	1.02×10^{-14}

Note that typically in a stochastic approximation framework step-sizes are kept constant in the beginning and then are slowly reduced. This is because initially the bias in the estimate is high and it reduces quickly if the step-sizes are large. However, large step sizes imply large noise or variance in the estimate so that once the bias has diminished the decreasing step-sizes help control the variance (See, e.g., Kushner and Yin [1997]). With this in mind, in our search for good step-sizes we restrict our attention to those of the form

$$\gamma_t(x) = \left\{ \begin{array}{ll} \pi & \text{if} \quad t(x) < N, \\ \pi(N/t(x))^{\alpha} & \text{otherwise,} \end{array} \right\},$$

where, as before, $t(x)$ denotes the number of iterations in the first t iterations in which state x is visited, and π, N and α are constants whose values are selected from experimentation (α is restricted to (0.5, 1] so that the conditions of Theorem A.1 on step-sizes hold). These experiments are shown in Tables I, II, and III. In each experiment the number of iterations (or transitions) per replication was set to 4×10^6 (it can be seen that the estimates corresponding to all the parameters considered are well within 1% of the true value in these many iterations). It was observed that on increasing the value of N beyond 50 the performance became worse; hence we restricted our attention to $N \leq 50$ in these experiments. Observe that $\pi = 0.5$, $N = 10$ and $\alpha = 0.8$ lead to the best performance. These are chosen in the remaining experiments with the ITDIS methods with Example 3.3.

Using the selected step-sizes we study the sensitivity of the algorithm to the value of λ. Table IV shows the MSE estimates for different values of λ from 0 to 0.81 in increments of 0.09. We see that $\lambda = 0.27$ in fact performs the best among the values of λ considered (though not reported, this observation was seen to be true even for other sets of step-sizes). We also observe that the performance deteriorates as λ gets closer to 1 or 0.

Table III. Estimated MSE for ITDIS Algorithm for Different Values of α and N at $\pi = 0.9$

		N		
		1	10	50
α	0.6	2.57×10^{-14}	5.11×10^{-14}	1.76×10^{-13}
	0.7	9.19×10^{-15}	3.87×10^{-14}	2.37×10^{-13}
	0.8	3.23×10^{-14}	2.21×10^{-14}	1.57×10^{-13}
	0.9	3.80×10^{-13}	9.93×10^{-15}	2.06×10^{-14}
	1.0	2.13×10^{-12}	8.30×10^{-15}	2.56×10^{-14}

Table IV. Estimated MSE for the ITDIS Algorithm for Different Values of λ

	0	0.09	0.18	0.27	0.36	0.45	0.54	0.63	0.72	0.81
MSE (10^{-14})	1.71	0.73	0.62	0.56	0.92	1.41	1.93	2.26	3.48	4.38

As mentioned earlier, empirically it has been observed that the best value of λ lies between 0 and 1. The value of λ that satisfies Lemma 4.2 in the examples we consider is typically small. Noting these facts and the results of Table IV, we use the highest value of λ that satisfies Lemma 4.2 in the remaining reported experiments with the ITDIS and the TDIS methods (unless otherwise stated).

As noted in Section 4.3, finding such a λ for general Markov chains may not be practical. In practice, one may choose to be conservative and select $\lambda = \min_{x \in \mathcal{I}, y \in \mathcal{S}} \frac{p_{xy}}{p_{xy}}$, if this is feasible (in Example 3.3, this equals 0.04). Again, $\lambda = 0$ guarantees that Theorem 4.3 holds, and as indicated in Table IV, empirically it performs better than naive simulation on this example (for the same computational effort, the MSE under naive simulation equaled 1.53×10^{-13}; see Figure 2). Thus, $\lambda = 0$ may be a reasonable choice in general Markov chain settings. Developing theoretical foundations for selecting good $(\gamma_t(\cdot) : t \geq 1)$, or in our settings π, N, and α, is an important problem that needs further research. Broadly speaking, π and/or N may be kept large when the bias of the initial guess $J_0(\cdot)$ is expected to be high. The parameter α may be kept high (close to 1) if the generated samples (e.g., $\sum_{m=n}^{\tau_t - 1} \lambda^{m-n} U_{m,t} \mathcal{L}_{n,m,t}$ in (15)) are noisy, that is, have large variance.

5.1.1 *Comparison of Algorithms.* We now compare the performance of the ITDIS algorithm with the TDIS, the IS, the NS, and the TD algorithms by estimating the MSE for each algorithm as a function of the computational budget per replication (as mentioned earlier, 20 independent replications of each method are conducted). Figure 2 plots the the estimated MSEs versus the computation effort for these methods. Figure 3 plots the estimated MSEs versus the computation effort for the ITDIS and the TDIS method on a magnified scale to highlight the improved performance of the ITDIS method. A unit of time in these figures corresponds roughly to 100,000 iterations of the ITDIS and the TDIS methods, 125,000 samples of the IS method, 600,000 iterations of the TD method, and 725,000 samples of the NS method. As with the ITDIS, the TDIS method uses step-sizes corresponding to $\pi = 0.5$, $N = 10$, and $\alpha = 0.8$ (though not reported, experiments conducted with the TDIS method for different step-sizes and λ showed that these parameters perform best for it as well).

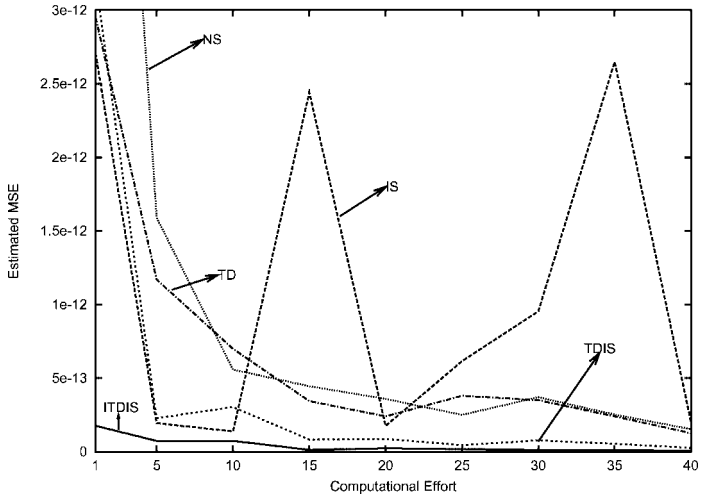

Fig. 2. Comparison of estimated MSE of ITDIS, TDIS, IS, TD, and NS methods. A unit of computational effort equals 3.75 seconds of computational time corresponding roughly to 100,000 iterations of the ITDIS method.

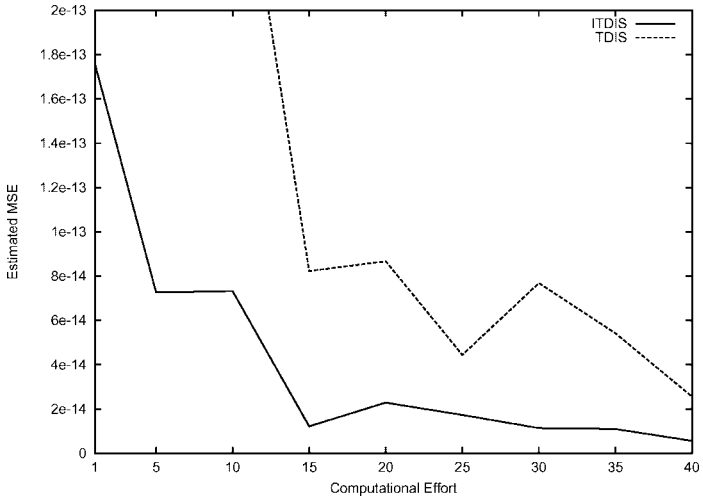

Fig. 3. Comparison of estimated MSE of ITDIS and TDIS methods on a magnified scale. A unit of computational effort equals 3.75 seconds of computational time corresponding roughly to 100,000 iterations of the ITDIS method.

Similarly, for the TD algorithm, we tried different values of step-size parameters and λ and then chose the ones that were seen to perform the best. These turned out to correspond to $\alpha = 1$, $\pi = 1$, $N = 1$, and $\lambda = 0.99$. In Section A.3 in the Appendix, we provide a heuristic justification for near optimality of these "extreme" parameters in our settings.

Figure 2 and 3 indicate that the ITDIS method performs the best, followed by the TDIS method. Recall from Section 4.1 that the ITDIS method always

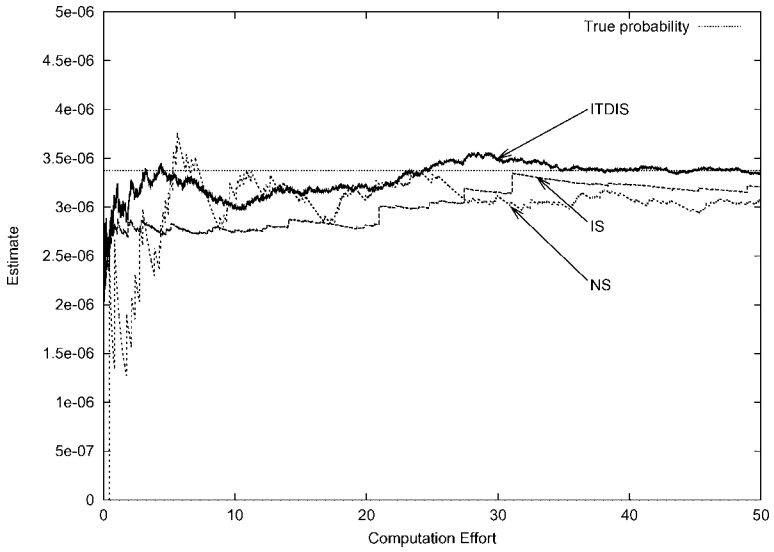

Fig. 4. A replication of ITDIS, IS, and NS methods. A unit of computational effort equals 3.75 seconds of computational time corresponding roughly to 100,000 iterations of the ITDIS method.

has nonnegative estimates while the TDIS method may even have negative estimates. This suggests that the ITDIS method is less noisy and has better convergence properties compared to the TDIS method, and this is reflected in the observed data. The TD method performs slightly better than the NS method. As expected, the MSE of the IS estimator behaves erratically (we discuss this further in the next paragraph).

Figure 4 visually illustrates the performance of the ITDIS, the IS, and the NS methods by plotting the estimate of the probability versus the computational effort for one replication of the algorithm (plots for the TDIS and the TD methods are not shown to facilitate clear presentation of the other three methods). From the plot, it appears that the ITDIS method performs much better than both the NS and the IS methods. Note that the plot of the IS estimator in Figure 4 shows the characteristic underestimation and large corrective jumps typical under many importance sampling schemes that have large variance (this example has infinite variance). The rationale for this is straightforward. Mostly the paths to the rare event that are assigned high probability under the new probability measure are observed. The likelihood ratio is small in such settings and hence the estimator typically underestimates when such samples predominate. Rarely, a corrective path to the rare event that is less likely under the new measure but more likely under the original measure is observed. Here, the likelihood ratio can be quite large, causing a jump in the estimate value (note that due to the law of large numbers even the IS estimator with infinite variance converges almost surely to the correct value). This also explains the erratic behavior of the observed MSE under the IS method in Figure 2. Recall that the theoretical MSE under the IS method is infinite. The estimated MSE reflects this by blowing up after a long time period when the corrective

samples appear. In a separate experiment, we ran 100 independent runs of the IS method each for 10 million iterations (80 computational effort units) and found that the estimated MSE further increased to about 2×10^{-12}.

Also, note that initially the plot of the ITDIS method in Figure 4 underestimates the actual value. This is due to the initial bias introduced by setting $(J_0(x) : x \in I)$ equal to zero. In Figure 4, the NS method shows high variance typical in rare event simulation. The fact that for large values of computational effort the NS method estimator underestimates the actual value in Figure 4 is a matter of chance. To see this heuristically, recall that a unit of computational effort in the graph corresponds to 725,000 samples of the busy cycle generated under the naive simulation. Thus, 30 units correspond to 2.175×10^7 samples. Since the number of samples is orders of magnitude higher than the reciprocal of the probability, it is reasonable to expect that the estimate is approximately normally distributed. Its mean and standard deviation can be seen to equal 3.37×10^{-6} and 0.39×10^{-6}, respectively. Also, note that the NS method estimator at 30 units of computational effort is approximately 3×10^{-6}. Thus, the observed value lies within 1 standard deviation of the actual value. The observed value after 50 computational effort units (3.625×10^7 samples) equals about 3.1×10^{-6}; thus the remaining 20 computational effort units (1.45×10^7 samples), the observed average, equaled about 3.25×10^{-6}. This, again, is not unusual as the mean and the standard deviation of this average are 3.37×10^{-6} and 0.48×10^{-6}, respectively. Given that the behavior of the NS method estimator at computational effort 30 and 50 is not unusual, the overall underestimation observed in the generated path is not atypical.

5.1.2 *Robustness of ITDIS Methods.* Recall that under the $P\&W$ change of measure, $(\nu_1', \nu_2', \mu_1', \mu_2') = (0.67, 0.07, 0.13, 0.13)$ and the new transition probabilities are $(b_{11}', b_{12}', b_{1e}') = (0, 0.26, 0.74)$, and $(b_{21}', b_{22}', b_{2e}') = (0.985, 0, 0.015)$. We now experimentally study the sensitivity of the ITDIS method with respect to the change of measure. In order to do so, we first vary the parameters ν_1', ν_2', μ_1', and μ_2' in a neighborhood of the $P\&W$ change of measure; maintaining their sum to be 1. Again, for each set of parameters chosen, we estimate the MSE generated by the ITDIS method. Computational effort of each replication is fixed at 150 seconds (this is the average time taken by the ITDIS method to complete 4×10^6 iterations with the $P\&W$ change of measure; this also equals 40 computational effort units in Figures 2, 3, and 4). Table V shows the values of the estimated MSEs. Observe that some of the changes of measures perform better than the $P\&W$ change of measure and all the changes of measures considered perform better than both the IS and the NS methods.

To check the sensitivity of the results to the transition probabilities, we also tested changes of measures whose transition probabilities differ from the $P\&W$ change of measure while the remaining rates are unchanged. Table VI displays the resultant performance. Again, note that the performance under the perturbed changes of measure is comparable to (though slightly worse than) that of the $P\&W$ change of measure.

From Theorem 4.3 it follows that the ITDIS method converges to the correct value for any change of measure if the λ is suitably selected. However, it does

Table V. Estimated MSE. (Recall that $\nu_1' = 1 - \mu_1' - \mu_2' - \nu_2'$. Also recall that under the $P\&W$ change of measure, $(\nu_1', \nu_2', \mu_1', \mu_2') = (0.67, 0.07, 0.13, 0.13)$. In these experiments $\alpha = 0.8$, $\pi = 0.5$, and $N = 10$. In each experiment λ is selected to be the highest value that satisfies Lemma 4.2)

		μ_2'		
μ_1'	ν_2'	0.10	0.13	0.16
0.10	0.04	8.36×10^{-14}	8.27×10^{-14}	6.89×10^{-14}
	0.07	3.49×10^{-14}	1.97×10^{-14}	2.52×10^{-14}
	0.10	1.33×10^{-14}	1.28×10^{-14}	1.36×10^{-14}
0.13	0.04	3.34×10^{-14}	3.02×10^{-14}	4.41×10^{-14}
	0.07	1.21×10^{-14}	5.61×10^{-15}	1.17×10^{-14}
	0.10	5.14×10^{-15}	5.73×10^{-15}	5.32×10^{-15}
0.16	0.04	1.61×10^{-14}	1.18×10^{-14}	2.38×10^{-14}
	0.07	4.17×10^{-15}	4.42×10^{-15}	5.77×10^{-15}
	0.10	2.35×10^{-15}	3.30×10^{-15}	2.98×10^{-15}

Table VI. Estimated MSE, $b_{11}' = b_{22}' = 0$. (Recall that under the $P\&W$ change of measure, $(b_{11}', b_{12}', b_{1e}') = (0, 0.26, 0.74)$, and $(b_{21}', b_{22}', b_{2e}') = (0.985, 0, 0.015)$. In these experiments $\alpha = 0.8$, $\pi = 0.5$, and $N = 10$. In each experiment λ is selected to be the highest value that satisfies Lemma 4.2)

	(b_{12}', b_{21}')		
	(0.20, 0.90)	(0.24, 0.985)	(0.30, 0.90)
MSE (10^{-14})	0.78	0.56	1.01

not throw light on the rate of convergence. The above experiments show that at least if the change of measure is close to the $P\&W$ change of measure we may expect a faster rate of convergence compared to naive simulation.

5.2 Second Example

We consider two queues in tandem with parameters $(\nu_1, \nu_2, \mu_1, \mu_2)$ equal to $(0.01, 0, 0.541, 0.449)$. The transition probabilities $(b_{11}, b_{12}, b_{1e}) = (0, 1, 0)$ and $(b_{21}, b_{22}, b_{2e}) = (0, 0, 1)$. Under the $P\&W$ change of measure the rates $(\nu_1', \nu_2', \mu_1', \mu_2') = (0.449, 0, 0.541, 0.01)$ and the transition probabilities remain unchanged, that is, $(b_{11}', b_{12}', b_{1e}') = (0, 1, 0)$ and $(b_{21}', b_{22}', b_{2e}') = (0, 0, 1)$. We select $B = 5$. For these parameters, the $P\&W$ change of measure has been proved to be asymptotically optimal in Glasserman and Kou [1995]. In this case, $P_{Overflow}$ was numerically found to equal 8.554×10^{-7}. With $P\&W$ change of measure, the ITDIS method was used with parameter $\lambda = 0.82$ (largest value that satisfies Lemma 4.2) and the step-size parameters $\alpha = 0.95$, $\pi = 1$, $N = 1$ (these performed best among a range of step-sizes tested). Table VII compares the estimated MSE of the estimate generated by the two methods. Each time unit corresponds to 0.21 seconds, the average time for 1000 iterations of the ITDIS method. Each iteration of the ITDIS method took on an average approximately 50% more time than the average time taken to generate a sample using the IS method. From the table we see that the ITDIS method performs marginally better than the IS method. This suggests that under all

Table VII. Comparison of the MSEs of ITDIS and IS Methods

	1	5	10	25	40
ITDIS	3.71×10^{-16}	8.00×10^{-17}	2.84×10^{-17}	1.18×10^{-17}	4.75×10^{-18}
IS	3.78×10^{-16}	6.78×10^{-17}	3.23×10^{-17}	1.35×10^{-17}	6.79×10^{-18}

sets of parameters the ITDIS method with suitably selected parameters may be expected to perform well.

The method of multiple independent replications that we used here to estimate MSEs can be used to construct confidence intervals and provide a stopping rule (e.g., stop when the 95% confidence interval is within 2% of the point estimate). However, it is wasteful as each replication may have a large initial bias (as indicated by Figure 2). Also, it is cumbersome when implementing a stopping rule as information from each of the replications needs to be stored. Alternatively, a single run and the method of batch means can be used to construct confidence intervals and effectively implement a stopping rule (see, e.g., Bratley et al. [1987] for a discussion on the method of multiple independent replications and the method of batch means). Also refer to Hseih and Glynn [2002] for another simple procedure to construct confidence intervals in the stochastic approximation setting.

6. POTENTIAL DRAWBACKS AND OPPORTUNITIES

In this paper we developed the combined temporal difference and importance sampling techniques in a finite state Markov chain framework and showed conditions under which the combined techniques converge to the correct value. A potential drawback of our approach is that we need to store the estimated value function for each state generated by the simulation. Note that storing these values is useful if the Markov chain under consideration results from a policy evaluation step in solving an MDP as these estimates can be used to come up with an improved policy. However, this storage requirement can be large if the state space of the Markov chain is huge (over hundreds of thousands of states) and if the paths generated via simulation visit a significant proportion of these states.

Also note that for some Markov chains it may be difficult to represent the states in a convenient manner so that the estimated value function may be stored for each state. For example, consider a single Markovian queue where multiple m classes of customers arrive. Here, for a given n number of customers, there are m^n states corresponding to different ordering of customers. The situation becomes much more complex if we consider a network of such queues. A convenient representation of states in Markov chains with such complex state descriptions becomes a problem. Thus, implementing numerical techniques and temporal difference-based simulation algorithms that require storing a value corresponding to each state may be difficult in such settings.

On the positive side, in the existing literature, temporal difference-based functional approximation techniques have been developed to approximately solve for performance measures related to Markov chains and Markov decision processes involving huge state spaces. These involve clever compact

representations of state space to reduce the storage requirements, although at the cost of some accuracy (see, e.g., Bertsekas and Tsitsiklis [1996]). We hope that this paper motivates further work to explore combining importance sampling with temporal difference methods using functional approximations to solve Markov chains and Markov decision processes involving huge state spaces.

APPENDIX

We first briefly review the Robbins-Monro stochastic approximation (SA) framework and state sufficient conditions under which the iterates from the algorithm are guaranteed to converge almost surely and their mean square error converges to zero. These are useful in proving Theorem 4.3.

A.1 Stochastic Approximation Framework

The stochastic approximation methods often involve solving for $J \in \mathcal{R}^n$ that satisfies a set of "fixed-point" equations $J = HJ$ where H is a mapping from $\mathcal{R}^n$ to itself (see, e.g., Kushner and Yin [1997]). In many situations, for a given $\tilde{J} \in \mathcal{R}^n$, it may be difficult to evaluate $H\tilde{J}$ while a noisy sample $H\tilde{J} + W$ may be readily available, where W is an n-dimensional random vector with zero mean. In such situations, to evaluate the fixed point, the algorithm proceeds in an iterative manner. The first iterate J_0 may be arbitrarily selected. If J_t denotes the iterate at time t then the update proceeds as follows:

$$J_{t+1} = (1 - \gamma_t) J_t + \gamma_t (H J_t + W_t), \tag{17}$$

where $(\gamma_t \in \Re^n : t \geq 0)$ is a sequence of appropriately chosen nonnegative step-sizes. The vectors in the sequence $(W_t \in \Re^n : t \geq 0)$ are uncorrelated zero mean random noise terms. As in Bertsekas and Tsitsiklis [1996], we focus on the following more general version of the above update method:

$$J_{t+1} = (1 - \gamma_t) J_t + \gamma_t (H_t J_t + W_t), \tag{18}$$

where each H_t may be stochastic (it may depend upon the history of the algorithm up till time t) and is assumed to belong to a family of mappings $\mathcal{H}$ with some desirable properties. Later in this section we show the explicit forms of H_t and W_t that allow representing the TDIS and the ITDIS algorithms as Equation (18).

Theorem A.1 states the sufficient conditions under which the above mentioned SA algorithm converges. The reader may refer, for example, to Bertsekas and Tsitsiklis [1996] and Borkar and Meyn [2000] for two distinct proofs of this theorem. Some notation is needed to help in its statement. Let $\mathcal{F}_t$ represent the history of the algorithm, that is, $\mathcal{F}_t = (J_0, J_1, \ldots, J_t, \gamma_1, \gamma_2, \ldots, \gamma_t, W_1, W_2, \ldots W_{t-1})$. For a vector $x = (x_1, \ldots, x_n) \in \Re^n$, the weighted maximum norm $\|x\|_\xi$ is defined as follows:

$$\|x\|_\xi = \max_{i \leq n} \frac{|x_i|}{\xi_i}, \tag{19}$$

where $\xi = (\xi_1, \ldots, \xi_n)$ is a vector in $\Re^n$ with positive components.

THEOREM A.1. *The sequence of random vectors* $(J_t : t \geq 1)$ *converges to* J *with probability* 1 *under the following conditions:*

(1) *For all* x, $\sum_{t=0}^{\infty} \gamma_t(x) = \infty$, *and* $\sum_{t=0}^{\infty} \gamma_t^2(x) < \infty$. *In case* $(\gamma_t : t \geq 0)$ *is a random sequence, these relations hold with probability* 1.

(2) *For every* x *and* t: $E[W_t(x)|\mathcal{F}_t] = 0$.

(3) *For some norm* $\|\cdot\|$ *on* $\mathcal{R}^n$ *there exist constants* A *and* B *such that*

$$E\left[W_t^2(x)|\mathcal{F}_t\right] \leq A + B\|J_t\|^2.$$

(4) *There exists a vector* J, *a positive vector* ξ, *and a scalar* $\beta \in [0, 1)$, *such that* $\|H_t\tilde{J} - J\|_\xi \leq \beta\|\tilde{J} - J\|_\xi$ *for all* $\tilde{J} \in \Re^n$.

Note that when J_t converges to J almost surely, it follows that $(J_t - J)^2$ converges to zero almost surely. Then, if for any $p > 2$,

$$\sup_t E[|J_t - J|^p] < \infty, \tag{20}$$

it follows that the sequence $((J_t - J)^2 : t \geq 0)$ is uniformly integrable and the MSE $E[(J_t - J)^2]$ converges to zero (see, e.g., Billingsley [1995]). The following lemma states the condition under which relation (20) holds. Its proof is essentially identical to the proof given in Borkar and Meyn [2000] when the two norm is replaced by the $p \geq 2$ norm and the constant step-size analysis is replaced by the decreasing step-size analysis. To avoid tedious repetition, the proof is omitted.

LEMMA A.2. *If, along with conditions* 1, 2 *and* 4 *in Theorem* A.1, *condition* 3 *is further strengthened so that there exist constants* A *and* B *and* $p > 2$ *such that for all* x,

$$E\left[W_t^p(x)|\mathcal{F}_t\right] \leq A + B\|J_t\|^p,$$

then relation (20) *holds.*

A.2 Proof of Theorem 4.3

As in Bertsekas and Tsitsiklis [1996], we first show that the update steps in both the TDIS and the ITDIS methods may be reexpressed in the form (18). We then note that the resulting $H_t(\cdot)$ is unchanged by the use of importance sampling and hence the proof given in Bertsekas and Tsitsiklis [1996] for step 4 of Theorem A.1 (the key step), where importance sampling is not used, holds for our case as well (the same is also true for step 1 of Theorem A.1). The remaining steps 2 and 3 of Theorem A.1 and Lemma A.2 are easily seen to be true in our setting.

We need to introduce new terminology to make our notation compatible with the notation used in Bertsekas and Tsitsiklis [1996] to facilitate using their results.

In the TDIS and the ITDIS algorithm, consider a generated sample path at iteration t. Suppose that $X_{n,t} = x$ for $n < \tau$ and that this is the first time the sample path visits state x. Let $Z_{m,t}(x) = 0$ for $m < n$ and $Z_{n,t}(x) = 1$. For $m > n$, let $Z_{m,t}(x) = \lambda Z_{m-1,t}(x) = \lambda^{m-n}$. Similarly, define $\mathcal{L}_{m,t}(x) = 1$ for $m < n$. Let

$\mathcal{L}_{m,t}(x) = L(X_{n,t}, \ldots, X_{m+1,t})$ for $m \geq n$. Then Equation (14) may be rewritten as

$$J_{t+1}(x) = J_t(x) + \gamma_t(x) \sum_{m=0}^{\tau_t - 1} Z_{m,t}(x) D_{m,t} \mathcal{L}_{m,t}(x). \tag{21}$$

Similarly, Equation (15) may be reexpressed as

$$J_{t+1}(x) = (1 - \gamma_t(x)) J_t(x) + \gamma_t(x) \left(\sum_{m=0}^{\tau_t - 1} Z_{m,t}(x) U_{m,t} \mathcal{L}_{m,t}(x) \right). \tag{22}$$

The $Z_{m,t}(x)$ are referred to as *eligibility coefficients* in the TD literature. The technique that we have used to assign values to each $Z_{m,t}(x)$ is referred to as the *First TD method*. Refer, for example, to Singh and Sutton [1996] and Singh and Dayan [1998] for other successful ways of assigning values to the eligibility coefficients. Our analysis easily extends to the more general TD methods. A key requirement is that the eligibility coefficients should decrease at least at a geometric rate so that a relation corresponding to relation (16) holds.

First consider the TDIS method. For notational convenience, we suppress the subscript $X_{0,t}$ from the operators E and $\bar{E}$ in the following discussion. Let $\delta_t(x)$ denote the probability of hitting state x in iteration t under the original probability measure.

Then, the update equation (21) may be reexpressed as

$$\begin{aligned} J_{t+1}(x) \;=\; & J_t(x)(1 - \gamma_t(x)\delta_t(x)) + \gamma_t(x)\delta_t(x) \\ & \times \left(\frac{\bar{E}\left[\sum_{m=0}^{\tau_t - 1} Z_{m,t}(x) \mathcal{L}_{m,t}(x) D_{m,t} | \mathcal{F}_t \right]}{\delta_t(x)} + J_t(x) \right) + \gamma_t(x)\delta_t(x) \\ & \times \left(\frac{\sum_{m=0}^{\tau_t - 1} Z_{m,t}(x) \mathcal{L}_{m,t}(x) D_{m,t} - \bar{E}\left[\sum_{m=0}^{\tau_t - 1} Z_{m,t}(x) \mathcal{L}_{m,t}(x) D_{m,t} | \mathcal{F}_t \right]}{\delta_t(x)} \right). \end{aligned}$$

Define mapping $H_t : \Re^n \mapsto \Re^n$ by letting for any vector $J \in \Re^n$

$$\begin{aligned} & (H_t J)(x) \\ & \quad = \frac{\bar{E}\left[\sum_{m=0}^{\tau_t - 1} Z_{m,t}(x) \mathcal{L}_{m,t}(x) (g(X_{m,t}, X_{m+1,t}) + J(X_{m+1,t}) - J(X_{m,t})) | \mathcal{F}_t \right]}{\delta_t(x)} \\ & \qquad + J(x). \end{aligned} \tag{23}$$

Note that

$$(H_t J_t)(x) = \frac{\bar{E}\left[\sum_{m=0}^{\tau_t - 1} Z_{m,t}(x) \mathcal{L}_{m,t}(x) D_{m,t} | \mathcal{F}_t \right]}{\delta_t(x)} + J_t(x). \tag{24}$$

By repeating the arguments used in the proof of Lemma 4.1, it follows that

$$(H_t J_t)(x) = \frac{E\left[\sum_{m=k}^{\tau_t - 1} Z_{m,t}(x) D_{m,t} | \mathcal{F}_t \right]}{\delta_t(x)} + J_t(x).$$

This, as mentioned earlier, is identical to the mapping used in Bertsekas and Tsitsiklis [1996]. Further note that the update step in the TDIS algorithm may

be reexpressed as

$$J_{t+1}(x) = (1 - \hat{\gamma}_t(x))\, J_t(x) + \hat{\gamma}_t(x)((H_t\, J_t)(x) + W_t(x)), \tag{25}$$

where

$$\hat{\gamma}_t(x) = \gamma_t(x)\delta_t(x) \tag{26}$$

(this satisfies step 1 in Theorem A.1 as $\delta_t(x)$ is a positive constant independent of t) and

$$W_t(x) = \frac{\sum_{m=0}^{\tau_t-1} Z_{m,t}(x)\mathcal{L}_{m,t}(x)D_{m,t} - \bar{E}\left[\sum_{m=0}^{\tau_t-1} Z_{m,t}(x)\mathcal{L}_{m,t}(x)D_{m,t}|\mathcal{F}_t\right]}{\delta_t(x)}. \tag{27}$$

Clearly step 2 of Theorem A.1 holds. We later show that step 3 and the condition for Lemma A.2 also hold.

We now focus on the ITDIS method. In this case, we may reexpress Equation (22) as

$$\begin{aligned} &J_{t+1}(x) \\ &= J_t(x)(1 - \hat{\gamma}_t(x)) + \hat{\gamma}_t(x)\left[\frac{\bar{E}\left[\sum_{m=0}^{\tau_t-1} Z_{m,t}(x)\mathcal{L}_{m,t}(x)U_{m,t}|\mathcal{F}_t\right] - J_t(x)}{\delta_t(x)} + J_t(x)\right] \\ &\quad + \hat{\gamma}_t(x)\left[\frac{\sum_{m=0}^{\tau_t-1} Z_{m,t}(x)\mathcal{L}_{m,t}(x)U_{m,t} - \bar{E}\left[\sum_{m=0}^{\tau_t-1} Z_{m,t}(x)\mathcal{L}_{m,t}(x)U_{m,t}|\mathcal{F}_t\right]}{\delta_t(x)}\right]. \end{aligned}$$

Note that $\bar{E}[\sum_{m=0}^{\tau_t-1} Z_{m,t}(x)\mathcal{L}_{m,t}(x)U_{m,t}|\mathcal{F}_t]$ equals $E[\sum_{m=0}^{\tau_t-1} Z_{m,t}(x)U_{m,t}|\mathcal{F}_t]$. In the TD method, by substituting for $U_{m,t}$ and $D_{m,t}$ it is easily seen that $E[\sum_{m=0}^{\tau_t-1} Z_{m,t}(x)U_{m,t}|\mathcal{F}_t] - J_t(x)$ equals $E[\sum_{m=0}^{\tau_t-1} Z_{m,t}(x)D_{m,t}|\mathcal{F}_t]$.

Therefore, the update step in ITDIS algorithm may be reexpressed as

$$J_{t+1}(x) = (1 - \hat{\gamma}_t(x))\, J_t(x) + \hat{\gamma}_t(x)((H_t\, J_t)(x) + \bar{W}_t(x)), \tag{28}$$

with H_t a mapping identical to that in the TDIS method and where

$$\bar{W}_t(x) = \left[\frac{\sum_{m=0}^{\tau_t-1} Z_{m,t}(x)\mathcal{L}_{m,t}(x)U_{m,t} - \bar{E}\left[\sum_{m=0}^{\tau_t-1} Z_{m,t}(x)\mathcal{L}_{m,t}(x)U_{m,t}|\mathcal{F}_t\right]}{\delta_t(x)}\right].$$

Again, step 2 of Theorem A.1 is easily seen to be true. If we show that there exist constants A and B such that, for all $x \in \mathcal{I}$,

$$E(W_t(x)^p|\mathcal{F}_t) \le A + B\|\, J_t\|^p \tag{29}$$

and

$$E(\bar{W}_t(x)^p|\mathcal{F}_t) \le A + B\|\, J_t\|^p \tag{30}$$

for all $p \ge 2$, then the proof of Theorem A.1 is complete. We show relation (29). Relation (30) can be similarly shown.

To prove relation (29), we focus on the numerator of $W_t(x)$ as the term $\delta_t(\cdot)$ is a deterministic term lower bounded by a positive constant for all $x \in \mathcal{I}$ and t. Using the triangle inequality, observe that

$$\bar{E}\left[\left\|\sum_{m=0}^{\tau_t-1} Z_{m,t}(x)\mathcal{L}_{m,t}(x)D_{m,t} - \bar{E}\left[\sum_{m=0}^{\tau_t-1} Z_{m,t}(x)\mathcal{L}_{m,t}(x)D_{m,t}|\mathcal{F}_t\right]\right\|^p |\mathcal{F}_t\right]$$

is less than or equal to

$$\bar{E}\left[\left(\left\|\sum_{m=0}^{\tau_t-1} Z_{m,t}(x)\mathcal{L}_{m,t}(x)D_{m,t}\right\|+\left\|\bar{E}\left[\sum_{m=0}^{\tau_t-1} Z_{m,t}(x)\mathcal{L}_{m,t}(x)D_{m,t}|\mathcal{F}_t\right]\right\|\right)^p|\mathcal{F}_t\right]. \tag{31}$$

Let G be an upper bound for $|g(x, y)|$. Then $D_{m,t} \leq 2\,\|J_t\| + G$. Note that from relation (16) it follows that $Z_{m,t}(x)\mathcal{L}_{m,t}(x) \leq C$ for all x. Thus,

$$\sum_{m=0}^{\tau_t-1} Z_{m,t}(x)\mathcal{L}_{m,t}(x) \leq C\tau_t, \tag{32}$$

and phrase (31) may be upper bounded by

$$\bar{E}[(C\tau_t(2\|J_t\| + G) + C\bar{E}(\tau_t)(2\|J_t\| + G))^p|\mathcal{F}_t].$$

In this equation using the fact that τ_t has finite moments of all orders and the fact that, for any nonnegative constants a and b,

$$(a + b)^p \leq 2^p(a^p + b^p),$$

the desired result follows.

A.3 Justification of Near Optimal Step-Sizes for TD Algorithm

It is reasonable to assume that the optimal step-sizes for $\lambda = 1$ are nearly equal to the optimal step-sizes for $\lambda = 0.99$. We now argue that $\alpha = 1$, $\pi = 1$, $N = 1$ are indeed near optimal for $\lambda = 1$. For $\lambda = 1$, Equation (5) reduces to

$$J_{t+1}(x) = (1 - \gamma_t(x))\,J_t(x) + \gamma_t(x)\sum_{m=n}^{\tau_t-1} g(X_{m,t}, X_{m+1,t}),$$

where $X_{n,t} = x$. Let $(K_i : i \geq 1)$ denote the i.i.d. samples of $\sum_{m=0}^{\tau_t-1} g(X_m, X_{m+1})$ with $X_0 = x$. Recall that $t(x)$ denotes the number of visits to state x till time t. Let $(\gamma_i(x) : i \leq t(x))$ denote the associated step-sizes. Then, supposing that at time t state x is visited, $J_{t+1}(x)$ has the same distribution as

$$\gamma_1(x)\left(\prod_{i=2}^{t(x)}(1 - \gamma_i(x))\right)K_1 + \gamma_2(x)\left(\prod_{i=3}^{t(x)}(1 - \gamma_i(x))\right)K_2 + \cdots + \gamma_{t(x)}(x)K_{t(x)}. \tag{33}$$

Note that the sum of the coefficients of $(K_i : i \leq t(x))$ equals

$$1 - \left(\prod_{i=1}^{t(x)}(1 - \gamma_i(x))\right),$$

which can be seen to be very close to 1 for large $t(x)$ for the step-sizes that we consider. It is easy to see that among all $(a_i : i \leq n)$ such that $\sum_{i\leq n} a_i = 1$, the variance of $\sum_{i\leq n} a_i K_i$ is minimized at $a_i = 1/n$ for all n. This suggests that near minimum variance is achieved for (33) when $\gamma_{t(x)}(x) = 1/t(x)$. This thus justifies the near optimality of $\alpha = 1$, $\pi = 1$, $N = 1$ when $\lambda = 0.99$.

Now we further heuristically argue that, in our rare event settings, it is reasonable to expect λ close to 1 to perform better than λ significantly less than 1 when all $(J_0(x) : x \in I)$ are initialized to zero. As an extreme case, consider $\lambda = 1$ and $\lambda = 0$. Note that rarely a sample path generated from the initial state $(0, 0)$ hits the common buffer B before returning to state $(0, 0)$. Once it does, the estimate of the value function of its immediate neighbor along the generated path (say $(1, 0)$) increases from zero to a positive value when $\lambda = 1$, while for $\lambda = 0$, it remains zero until many such paths to the rare event occur so that on a generated path the state that is visited after the state $(1, 0)$ (say x) has a positive estimate of the value function (this would happen if, on a previously generated path to the rare event that included x, the state visited after x had a positive estimate of the value function, and so on). This heuristically suggests that in our settings as events become rarer due to increase in B, the bias vanishes faster for $TD(\lambda)$ for λ close to 1 and hence better performance may be expected for such λ's.

A.4 Parekh and Walrand Change of Measure

Extend the notation from Section 3.2 for two-queue Jackson network to a Jackson network with d queues (with single server at each queue). We assume that the network is stable, that is, $\gamma_i < \mu_i$ for all i. We further assume without loss of generality that queue 1 has the largest load, that is, $\rho_1 > \rho_i$ for $i > 1$, where $\rho_i = \gamma_i/\mu_i$. For each i, let r_i denote the the expected number of visits to queue 1 by a customer entering the network at queue i. Denote the $P\&W$ change of measure by appending a prime (" $'$ ") to the original notation. The following equation, derived by Frater et al. [1991], determines this change of measure:

$$\gamma_i' = \gamma_i \left[1 + \frac{r_i(\mu_1 - \gamma_1)}{r_1 \gamma_1} \right].$$

This implies that $\gamma_1' = \mu_1$. Also, $\nu_i' = \nu_i \frac{\gamma_i'}{\gamma_i}$,

$$\mu_1' = \gamma_1 + \frac{(r_1 - 1)(\mu_1 - \gamma_1)}{r_1},$$

and for all $i > 1$, $\mu_i' = \mu_i$. The new transition probabilities are given as $b_{ie}' = b_{ie} \frac{\gamma_i}{\min(\gamma_i', \mu_i')}$, for all i, and $b_{ij}' = b_{ij} \frac{\gamma_i}{\min(\gamma_i', \mu_i')} \frac{\gamma_j'}{\gamma_j}$ for all i and j.

ACKNOWLEDGMENTS

The authors would like to thank the journal area editor, the journal associate editor, and the two referees for their input, which led to considerable improvement in the paper.

REFERENCES

ANDRADOTTIR, S., HEYMAN, D., AND OTT, T. J. 1993. Variance reduction through smoothing and control variates for Markov chain simulations. *ACM Trans. Mod. Comput. Sim. 3*, 3, 167–189.

ANDRADOTTIR, S., HEYMAN, D., AND OTT, T. J. 1995. On the choice of alternative measures in importance sampling with Markov chains. *Operat. Res. 43*, 3, 509–519.

BERTSEKAS, D. AND TSITSIKLIS, J. 1996. *Neuro-Dynamic Programming*. Athena Scientific, Belmont, MA.

BILLINGSLEY, P. 1995. *Probability and Measure*. John Wiley & Sons, New York, NY.

BORKAR, V. AND MEYN, S. 2000. The o.d.e. method for convergence of stochastic approximation and reinforcement learning. *SIAM J. Contr. Opt. 38*, 2, 347–469.

BRATLEY, P., FOX, B., AND SCHRAGE, L. 1987. *A Guide to Simulation*. Springer-Verlag, New York, NY.

CRANE, M. AND IGLEHART, D. 1975. Simulating stable stochastic systems, iii: Regenerative processes and discrete-event simulations. *Operat. Res. 23*, 33–45.

DEMBO, A. AND ZEITOUNI, O. 1992. *Large Deviations Techniques and Applications*. Jones and Bartlett, Boston, MA.

FRATER, M., LENNON, T., AND ANDERSON, B. 1991. Optimally efficient estimation of the statistics of rare events in queuing networks. *IEEE Trans. Automat. Contr. 36*, 12, 1395–1404.

GLASSERMAN, P. AND KOU, S. 1995. Analysis of an importance sampling estimator for tandem queues. *ACM Trans. Mod. Comput. Sim. 5*, 1, 22–42.

GLASSERMAN, P. AND WANG, Y. 1997. Counterexamples in importance sampling for large deviations probabilities. *Ann. Appl. Prob. 7*, 731–746.

GLYNN, P. AND IGLEHART, D. 1989. Importance sampling for stochastic simulations. *Manage. Sci. 35*, 11, 1367–1392.

HEIDELBERGER, P. 1977. Variance reduction techniques for simulating Markov chains. In *Proceedings of the 1977 Winter Simulation Conference*.

HEIDELBERGER, P. 1980a. Variance reduction techniques for the simulation of Markov processes, i: Multiple estimates. *IBM J. Res. Develop. 24*, 570–581.

HEIDELBERGER, P. 1980b. Variance reduction techniques for the simulation of Markov processes, ii: Matrix iterative methods. *Acta Informatica 13*, 21–37.

HEIDELBERGER, P. 1995. Fast simulation of rare events in queuing and reliability models. *ACM Trans. Model. Comput. Sim. 5*, 1, 43–85.

HSEIH, M. AND GLYNN, P. 2002. Confidence regions for stochastic approximation algorithms. In *Proceedings of the 2002 Winter Simulation Conference*. 370–376.

JUNEJA, S. 2001. Importance sampling and the cyclic approach. *Operat. Res. 49*, 6, 900–912.

JUNEJA, S. 2003. Efficient rare event simulation using importance sampling: An introduction. In *Computational Mathematics, Modelling and Algorithms*, J. C. Misra, Ed. Narosa Publishing House, New Delhi, India, 357–396.

JUNEJA, S. AND SHAHABUDDIN, P. 2001. Efficient simulation of Markov chains with small transition probabilities. *Manage. Sci. 47*, 4, 547–562.

KUSHNER, H. AND YIN, G. 1997. *Stochastic Approximation Algorithms and Applications*. Springer-Verlag, New York, NY.

PAREKH, S. AND WALRAND, J. 1989. A quick simulation method for excessive backlogs in networks of queues. *IEEE Trans. Automat. Contr. 34*, 1, 54–66.

PRECUP, D., SUTTON, R., AND SINGH, S. 2000. Eligibility traces for off-policy policy evaluation. In *Proceedings of the 17th International Conference on Machine Learning (ICML'00)*. Morgan Kaufman, San Francisco, CA, 759–766.

RUBINSTEIN, R. 1997. Optimization of computer simulation models with rare events. *European J. Operat. Res. 99*, 89–112.

RUBINSTEIN, R. 1999. Rare event simulation via cross-entropy and importance sampling. *Second Workshop on Rare Event Simulation* (RESIM'99). 1–17.

SINGH, S. AND DAYAN, P. 1998. Analytical mean squared error curves for temporal difference learning. *Mach. Learn. 32*, 5–40.

SINGH, S. AND SUTTON, R. 1996. Reinforcement learning with replacing eligibility traces. *Mach. Learn. 22*, 123–158.

SUTTON, R. 1988. Learning to predict by the methods of temporal differences. *Mach. Learn. 3*, 9–44.

SUTTON, R. AND BARTO, A. 1998. *Reinforcement Learning: An Introduction*. The MIT Press, Cambridge, MA.

Received March 2002; revised February 2003, September 2003, October 2003; accepted October 2003