Mathematical Methods of Operations Research (1997) 46:241–250

On Confidence Intervals from Simulation of Finite Markov Chains

Apostolos N. Burnetas

Department of Operations Research and Operations Management, Weatherhead School of Management, Case Western Reserve University, 10900 Euclid Ave., Cleveland, OH 44106, USA
e-mail: atb4@po.cwru.edu

Michael N. Katehakis

Faculty of Management and RUTCOR, Rutgers University, 92 New Street, Newark, NJ 07102, USA
e-mail: mnk@andromeda.rutgers.edu

Abstract: Consider a finite state irreducible Markov reward chain. It is shown that there exist simulation estimates and confidence intervals for the expected first passage times and rewards as well as the expected average reward, with 100% coverage probability. The length of the confidence intervals converges to zero with probability one as the sample size increases; it also satisfies a large deviations property.

Key Words: Discrete Markov Chains, Simulation.

1 Introduction

Consider a finite state irreducible Markov chain, endowed with a reward structure. Simulation is often used for the estimation of quantities of interest such as the expected long-run average reward, expected first passage rewards, and the second moment of first passage rewards (c.f. Kleijnen (1992), Fishman (1978) and references therein). In this paper we show that there exist random intervals, generated by simulation estimates, for the above quantities, with the following properties: (a) they contain the respective quantities with probability one (sample-pathwise) and (b) they converge to the corresponding quantities with probability one, i.e., they are 100% simultaneous confidence intervals, with length converging to zero.

The derivation of the upper and lower bounds specifying the confidence intervals is based on the fact that the estimation errors satisfy equations of the same type as the estimated quantities.

In this paper we present the main idea of the method and convergence properties. Generalizations and adaptation of the method for efficient implementation deserve further study.

1432–2994/97/46:2/241–250 $2.50 © 1997 Physica-Verlag, Heidelberg

Bounds for the error in the average and discounted rewards for a Markov process subject to small perturbations of the transition matrix are developed in Van-Dijk & Puterman (1988). The ideas of the present paper have been employed in Burnetas & Katehakis (1996), for calculating average-optimal policies for Markovian Decision Processes using simulation. Related results, in the context of Markovian Decision Processes under incomplete information, are contained among others in Federgruen & Schweitzer (1981), Hernández-Lerma (1989), Burnetas & Katehakis (1997) and Katehakis & Robbins (1995), where the estimation of the optimal expected average or discounted rewards is performed via adaptive estimation of the transition probabilities, which are assummed to be only partially known. For further recent work we refer to Fishman (1994), for Markov Chain sampling, and Glasserman & Liu (1996), for simulation of multistage production systems.

In the next section we give a necessary background. In section 3 we present the details of the estimation scheme and prove the consistency of the estimators. In sections 4 and 5 we show the existence of 100% confidence intervals. In section 6 we prove a large deviation property related to the rate of decrease of the confidence intervals length.

2 Background

Consider a finite-state, positive recurrent Markov Reward Process $\{X_t, t = 0, 1, \ldots\}$, with state space $S = \{0, 1, \ldots, N\}$, transition matrix $P = [p_{xy}, x, y \in S]$ and reward vector $r = [r(x), x \in S]$. Let $\mathbf{P}_x$ and $\mathbf{E}_x$ denote probability and expectation given $X_0 = x$.

Let $\beta_0 = 0$ and $\beta_k = \min\{n \geq 1 : X_n = 0, X_t \neq 0, t = \beta_{k-1} + 1, \ldots, n-1\}$, $k = 1, 2, \ldots$, denote the successive return epochs to a reference state 0.

Define $m(x) = \mathbf{E}_x\beta_1$, $w(x) = \mathbf{E}_x \sum_{t=0}^{\beta_1 - 1} r(X_t)$ and $s(x) = \mathbf{E}_x(\sum_{t=0}^{\beta_1-1} r(X_t))^2$ as the the expected first passage time, expected first passage reward and the second moment of the first passage reward, respectively, from state x to state 0.

Also let $g = \lim_{T\to\infty} \mathbf{E}_x \sum_{t=0}^{T} r(X_t)/(T+1) = w(0)/m(0)$ denote the expected long run average reward.

It is well known that $m(x), w(x), s(x)$, $x \in S$ are unique solutions to systems of linear equations, c.f. Hordijk (1974),

$$m(x) = 1 + \sum_{y\neq 0} p_{xy} m(y) \ , \quad x \in S \tag{1}$$

$$w(x) = r(x) + \sum_{y\neq 0} p_{xy} w(y) \ , \quad x \in S \tag{2}$$

$$s(x) = 2r(x)w(x) - r^2(x) + \sum_{y\neq 0} p_{xy} s(y) \ , \quad x \in S \ . \tag{3}$$

In our notation $m(0)$ and $w(0)$ represent the expected number of steps and the expected reward between two successive visits to state 0. Therefore, the summations on the right hand side must explicitly exclude the term for $y = 0$.

In addition, (c.f. Derman (1970))

$$g + h(x) = r(x) + \sum_{y \in S} p_{xy} h(y) , \quad x \in S , \tag{4}$$

where h is a function on S defined up to an additive constant. If the normalization $h(0) = 0$ is adopted, then $h(x)$ can be interpreted as the expected first passage differential reward from x to 0, i.e., $h(x) = \mathbf{E}_x \sum_{t=0}^{\beta_1 - 1} (r(X_t) - g) = w(x) - gm(x)$.

Remark 1: In the remainder of the paper we assume, without loss of generality, that $r(x) > 0$, $\forall x \in S$. Indeed, if some of the rewards are negative, consider a modified problem with the same state space and transition mechanism, and rewards $r'(x) = r(x) + c > 0$, where $c > -\min_{x \in S} r(x)$. The quantities of interest for the initial and modified problems are related as follows; $m(x) = m'(x)$, $w(x) = w'(x) - cm(x), g = g' - c$. Therefore, any bounds developed for m', w', g' can be extended to the general case.

3 Estimation Procedures

Let a *cycle* denote the time interval between successive visits to state 0. A cycle constitutes a sample in our estimation procedure, and the terms cycle and sample will be used interchangeably.

We define the following random variables on the space of sample paths: $A_k(x) = \min\{t : \beta_k \le t \le \beta_{k+1} - 1, X_t = x\}$, $I_k(x) = \mathbf{1}\{A_k(x) < \infty\}$, $T_k(x) = I_k(x)(\beta_{k+1} - A_k(x))$, $W_k(x) = I_k(x) \sum_{t=A_k(x)}^{\beta_{k+1}-1} r(X_t)$ and $S_k(x) = I_k(x) \times (\sum_{t=A_k(x)}^{\beta_{k+1}-1} r(X_t))^2$.

Observe that $A_k(x)$ represents the epoch of first visit to state x during the k^{th}cycle, where $\min \emptyset = +\infty$. Similarly, $I_k(x)$ is the indicator of the event that state x is visited during the k^{th}cycle, $T_k(x)$ the time between first visit to x in the k^{th}cycle and the $(k + 1)^{st}$ recurrence to 0, (or zero if such a visit does not happen), and $W_k(x), S_k(x)$ the total reward obtained during the same period and its square, respectively.

The main idea of regenerative simulation is that for $k = 1, 2, \ldots$ the random vectors

$$V_k = \{\beta_{k+1} - \beta_k, X_t, t = \beta_k + 1, \ldots, \beta_{k+1}\} , \quad x \in S , \quad k = 1, 2, \ldots$$

are independent and identically distributed (cf. Hordijk, Iglehart & Schassberger (1976), Fishman (1978)). Therefore, $\forall x \in S$, $\{I_k(x), k = 1, 2, \ldots\}$, $\{T_k(x), k = 1, 2, \ldots\}$, $\{W_k(x), k = 1, 2, \ldots\}$ and $\{S_k(x), k = 1, 2, \ldots\}$ are sequences of i.i.d. random variables. Let **P** and **E** denote probability and expectation, respectively, with respect to the common distribution of the random vectors V_k.

Lemma 1:

1. $m(x) = \mathbf{E}[T_k(x)]/\mathbf{P}[I_k(x) = 1]$.
2. $w(x) = \mathbf{E}[W_k(x)]/\mathbf{P}[I_k(x) = 1]$.
3. $g = \mathbf{E}[W_k(0)]/\mathbf{E}[T_k(0)]$.
4. $s(x) = \mathbf{E}[S_k(x)]/\mathbf{P}[I_k(x) = 1]$.

Proof: By definition of $T_k(x)$, $\mathbf{E}[T_k(x)] = \mathbf{P}[I_k(x) = 1]\mathbf{E}[T_k(x)|I_k(x) = 1]$. Since all states are positive recurrent, $\mathbf{P}[I_k(x) = 1] > 0$, $x \in S$. In addition, $\mathbf{E}[T_k(x)I_k(x) = 1] = \mathbf{E}[T_1(x)|I_1(x) = 1] = \mathbf{E}_x[\beta_1] = m(x)$, where the first equality follows from the fact that $T_k(x), k = 1, 2, \ldots$ are i.i.d., and the second from the Markov property. This proves part 1.

The proof of parts 2 and 4 is similar. For 3, recall that $g = w(0)/m(0)$. □

For $n \geq 1$, $x \in S$, define the following estimators: $\hat{m}_n(x) = \bar{T}_n(x)/\bar{I}_n(x)$, $\hat{w}_n(x) = \bar{W}_n(x)/\bar{I}_n(x)$, $\hat{s}_n(x) = \bar{S}_n(x)/\bar{I}_n(x)$, $\hat{g}_n = \hat{w}_n(0)/\hat{m}_n(0)$ and $\hat{h}_n(x) = \hat{w}_n(x) - \hat{g}_n\hat{m}_n(x)$, where $\bar{I}_n, \bar{T}_n, \bar{W}_n, \bar{S}_n$ denote sample averages, i.e., $\bar{I}_n = 1/n \sum_{k=1}^{n} I_k(x)$, etc.

The estimators $\hat{m}_n(x), \hat{w}_n(x), \hat{s}_n(x)$ are well defined when $\bar{I}_n(x) > 0$. If this is not true, let $\hat{m}_n(x), \hat{w}_n(x)$ and $\hat{s}_n(x)$ be arbitrary positive numbers.

The next result follows from the strong law of large numbers and Lemma 1.

Proposition 1: The quantities $\hat{m}_n(x)$, $\hat{w}_n(x)$, $\hat{s}_n(x)$, $\hat{g}_n$, $\hat{h}_n(x)$ are strongly consistent estimators of $m(x)$, $w(x)$, $s(x)$, g and $h(x)$, respectively.

4 Expected Average Reward

In this section we derive a confidence interval with coverage probability equal to one for the expected long run average reward g. This is accomplished by developing upper and lower bounds for the estimation error based on a sample of size n.

Let $\delta^g_n = \hat{g}_n - g$, $\delta^h_n(x) = \hat{h}_n(x) - h(x)$ and $\Delta^g_n(x) = \hat{g}_n + \hat{h}_n(x)_r(x) - \sum_{y \in S} p_{xy}\hat{h}_n(y)$, $x \in S$, denote respectively the estimation error of $\hat{g}_n, \hat{h}_n(x)$ and the deviations between the left and right hand sides of the system equations (4), when the estimates for $g, h(x)$ are used instead of the true values. Also, let $\overline{\Delta}^g_n = \max_{x \in S}\Delta^g_n(x)$ and $\underline{\Delta}^g_n = \min_{x \in S}\Delta^g_n(x)$.

Proposition 2:

1. *An 100% confidence interval for g is $\hat{g}_n - U^g_n \leq g \leq \hat{g}_n - L^g_n$, where $L^g_n = \underline{\Delta}^g_n$ and $U^g_n = \min\{\hat{g}_n, \overline{\Delta}^g_n\}$.*
2. *Let $E^g_n = U^g_n - L^g_n$ denote the length of the confidence interval for g. Then $E^g_n \to 0$, with probability one as $n \to \infty$.*

Proof: Substituting δ^g_n and $\delta^h_n(x)$ into $\Delta^g_n(x)$ and using (4),

$$\Delta^g_n + \delta^h_n(x) = \Delta^g_n(x) + \sum_{y \neq 0} p_{xy}\delta^h_n(y).$$

Therefore, δ^g_n represents the expected average reward of a Markov Reward Process $\{Y_t, t \geq 0\}$, with transition matrix P and reward in state x equal to $\Delta^g_n(x)$. Hence, $\delta^g_n = \sum_{x \in S}\pi(x)\Delta^g_n(x)$, where π is the steady state probability vector of P.

It follows that $\underline{\Delta}^g_n \leq \delta^g_n \leq \overline{\Delta}^g_n$. In addition, $\delta^g_n < \hat{g}_n$, because $g = \hat{g}_n - \delta^g_n > 0$. Therefore, $L^g_n \leq \delta^g_n \leq U^g_n$, from which part 1 follows.

Part 2 is a consequence of part 1 and Proposition 1. □

Remark 2: From Proposition 2 it follows after simple algebra that the relative estimation error δ^g_n/g satisfies $L^g_n/(\hat{g}_n - L^g_n) \leq \delta^g_n/g \leq U^g_n(\hat{g}_n - U^g_n)$.

5 Expected First Passage Times and Rewards

In this section we develop 100% confidence intervals for $m(x), w(x)$ and $s(x), x \in S$. The proofs are similar to the proof of Proposition 2, using the appropriate interpretations of the estimation errors for each case.

Let $\delta^m_n(x) = \hat{m}_n(x) - m(x)$ and $\Delta^m_n(x) = \hat{m}_n(x) - 1 - \sum_{y \neq 0} p_{xy}\hat{m}_n(y)$, $x \in S$. Also let $\underline{\Delta}^m_n = \min_{x \in S}\Delta^m_n(x)$, $\overline{\Delta}^m_n = \max_{x \in S}\Delta^m_n(x)$.

For $w(x)$ define similarly, $\delta_n^w(x) = \hat{w}_n(x) - w(x)$, $\Delta_n^w(x) = \hat{w}_n(x) - r(x) - \sum_{y \neq 0} p_{xy}\hat{w}_n(y)$ and $\underline{\Delta}_n^m = \min_{x \in S} \Delta_n^w(x), \overline{\Delta}_n^w = \max_{x \in S} \Delta_n^w(x)$.

For $s(x)$, $\delta_n^s(x) = \hat{s}_n(x) - s(x)$, $\Delta_n^s(x) = \hat{s}_n(x) - 2r(x)\hat{w}_n(x) + r^2(x) - \sum_{y \neq 0} p_{xy}\hat{s}_n(y)$.

Proposition 3:

1. *An 100% confidence interval for $m(x)$ is*

$$\hat{m}_n(x) - U_n^m(x) \leq m(x) \leq \hat{m}_n(x) - L_n^m(x) ,$$

where

$$L_n^m(x) = l_n^m \hat{m}_n(x)/(1 + l_n^m) ,$$

$$U_n^m(x) = u_n^m \hat{m}_n(x)/(1 + u_n^m) ,$$

$$l_n^m = \max\{\underline{\Delta}_n^m, -1\} \textit{ and } u_n^m = \overline{\Delta}_n^m .$$

2. *An 100% confidence interval for $w(x)$ is*

$$\hat{w}_n(x) - U_n^w(x) \leq w(x) \leq \hat{w}_n(x) - L_n^w(x) ,$$

where

$$L_n^w(x) = \underline{\Delta}_n^w \hat{m}_n(x)/\rho_l^w ,$$

$$U_n^w(x) = \min\{\hat{w}_n(x), \overline{\Delta}_n^w \hat{m}_n(x)/\rho_u^w\} ,$$

$$\rho_l^w = \begin{cases} 1 + l_n^m , & \textit{if } \underline{\Delta}_n^w < 0 \\ 1 + u_n^m , & \textit{if } \underline{\Delta}_n^w \geq 0 \end{cases}$$

$$\rho_u^w = \begin{cases} 1 + u_n^m , & \textit{if } \underline{\Delta}_n^w < 0 \\ 1 + l_n^m , & \textit{if } \underline{\Delta}_n^w \geq 0 \end{cases}$$

3. *An 100% confidence interval for $s(x)$ is*

$$\hat{s}_n(x) - U_n^s(x) \leq s(x) \leq \hat{s}_n(x) - L_n^s(x),$$

where

$$L_n^s(x) = \underline{\Delta}_n^s \hat{m}_n(x)/\rho_l^s \ ,$$

$$U_n^s(x) = \min\{\hat{s}_n(x), \overline{\Delta}_n^s \hat{m}_n(x)/\rho_u^s\} \ ,$$

$$\rho_l^s = \begin{cases} 1 + l_n^m \ , & \textit{if } \underline{\Delta}_n^s < 0 \\ 1 + u_n^m \ , & \textit{if } \underline{\Delta}_n^s \geq 0 \end{cases}$$

$$\rho_u^s = \begin{cases} 1 + u_n^m \ , & \textit{if } \overline{\Delta}_n^s < 0 \\ 1 + l_n^m \ , & \textit{if } \overline{\Delta}_n^s \geq 0 \end{cases}$$

$$\underline{\Delta}_n^s = \min_{x \in S}\{2r(x)L_n^w(x) + \Delta_n^s(x)\} \ ,$$

$$\overline{\Delta}_n^s = \max_{x \in S}\{2r(x)U_n^w(x) + \Delta_n^s(x)\} \ .$$

4. *Let $E_n^j(x) = U_n^j(x) - L_n^j(x), j = m, w, s$ denote the length of the confidence intervals for $m(x), w(x), s(x)$ respectively. Then $E_n^j(x) \to 0$, with probability one as $n \to \infty$.*

Proof: Substituting $\delta_n^m(x)$ into $\Delta_n^m(x)$ and using (1),

$$\delta_n^m(x) = \Delta_n^m(x) + \sum_{y \neq 0} p_{xy}\delta_n^m(y) \ .$$

Thus $\delta_n^m(x)$ represents the expected first passage reward from x to 0 for a Markov Reward Process $\{Y_t, t \geq 0\}$, with transition matrix P and reward in state x equal to $\Delta_n^m(x)$. Therefore, $\delta_n^m(x) = \mathbf{E}\left[\sum_{t=0}^{\beta_1 - 1} \Delta_n^m(Y_t) | Y_0 = x\right]$ and

$$m(x)\underline{\Delta}_n^m \leq \delta_n^m(x) \leq m(x)\overline{\Delta}_n^m \ .$$

Note also that, by definition, $\hat{m}_n(x) > 0$, thus, $\delta_n^m(x) > -m(x)$.

Combining these two inequalities, it follows after some algebra that $L_n^m(x) \leq \delta_n^m(x) \leq U_n^m(x)$. This proves part 1.

Parts 2 and can be shown in a similar manner, using equations (2) and (3) respectively, as well as the results of part 1, for 2, and parts 1 and 2, for 3.

Part 4 follows from the expressions for the bounds and Proposition 1. □

Remark 3: (a) Using Proposition 3, the following bounds for the relative estimation errors are easily derived:

$$l_n^m \leq \frac{\delta_n^m(x)}{m(x)} \leq u_n^m ,$$

$$\frac{L_n^w(x)}{\hat{w}_n(x) - L_n^w(x)} \leq \frac{\delta_n^w(x)}{w(x)} \leq \frac{U_n^w(x)}{\hat{w}_n(x) - U_n^w(x)} ,$$

$$\frac{L_n^s(x)}{\hat{s}_n(x) - L_n^s(x)} \leq \frac{\delta_n^s(x)}{s(x)} \leq \frac{U_n^s(x)}{\hat{s}_n(x) - U_n^s(x)} .$$

(b) Due to the strong consistency of $\hat{m}_n(x)$, both $\delta_n^m(x) = \hat{m}_n(x) - m(x) \to 0$ and $\Delta_n^m(x) \to 0$ with probability one, as the sample size increases. Thus $1 + l_n^m > 0$ with probability one for large number of cycles, therefore the bounds presented in Proposition 3 are not trivial.
(c) By setting $r(x) = 1$, $\forall x \in S$, the results of Proposition 3 yield confidence intervals for the second moments of the first passage times.
(d) Using the bounds for the first and second moments of the first passage times and rewards, it is easy to develop bounds for the corresponding variances.

6 Rate of Convergence

Using the strong consistency of the estimators it was shown in sections 4 and 5 that the length of the derived confidence intervals decreases to zero with probability one.

In this section we show (Proposition 4) that, as a consequence of large deviations properties of the estimators, the probabilities $\mathbf{P}[E_n^j > \varepsilon]$, $j = g, m, w, s$, vanish exponentially with n, for all $\varepsilon > 0$.

This is equivalent to the following statement for the rate of of decrease of E_n^j. For any $\varepsilon >, \delta > 0$ there exists a $n_0 = n_0(\delta, \varepsilon) = O(|\log \delta|)$, $\forall \varepsilon > 0$, such that $\mathbf{P}[E_n^j > \varepsilon] \leq \delta$, $\forall n \geq n_0$.

Let $p(x) = \mathbf{E}[I_1(x)] = \mathbf{P}[I_1(x) = 1]$.

Lemma 2:

1. $\forall x \in S$, $\forall \varepsilon > 0$, $\exists \gamma^I = \gamma^I(x, \varepsilon) > 0$, *such that*

$$\mathbf{P}[|\bar{I}_n(x) - p(x)| > \varepsilon] \leq 2e^{-\gamma^I n} , \quad \forall n \geq 1 .$$

2. $\forall x \in S$, $\forall \varepsilon > 0$, $\exists \gamma^T = \gamma^T(x, \varepsilon) > 0$, *such that*

$$\mathbf{P}[|\bar{T}_n(x) - m(x)p(x)| > \varepsilon] \leq 2e^{-\gamma^T n} , \quad \forall n \geq 1 .$$

3. $\forall x \in S$, $\forall \varepsilon > 0$, $\exists \gamma^W = \gamma^W(x, \varepsilon) > 0$, *such that*

$$\mathbf{P}[|\overline{W}_n(x) - w(x)p(x)| > \varepsilon] \leq 2e^{-\gamma^W n}, \quad \forall n \geq 1 .$$

Proof: Fix $x \in S$. Let $\Lambda_{I(x)}(\theta) = \log \mathrm{E}[e^{\theta I_1(x)}]$ be the logarithm of the moment generating function of $I_1(x)$ and $\Lambda^*_{I(x)}(z) = \sup_{\theta \in \mathfrak{R}}[\theta z - \Lambda_{I(x)}(\theta)]$ the Legendre-Fenchl transform of $\Lambda_{I(x)}$. Note that $\Lambda^*_{I(x)}(p(x)) = 0$ and $\Lambda^*_{I(x)}(z) > 0$, $\forall z \neq p(x)$.

Then it follows from standard results of large deviations theory (c.f. Dembo & Zeitouni (1993)) that $\mathbf{P}[|\overline{I}_n(x) - p(x)| > \varepsilon] \leq 2e^{-\gamma^I n}$, where $\gamma^I = \gamma^I(x, \varepsilon) = \min(\Lambda^*_{I(x)}(p(x) - \varepsilon), \Lambda^*_{I(x)}(p(x) + \varepsilon))$. This proves part 1. The proof of parts 2 and 3 in similar. □

Lemma 3: Let $\mathbf{Z}, \mathbf{Z}^1, \ldots \mathbf{Z}^n$ *be i.i.d. random vectors* $\mathbf{Z} = (Z_1, \ldots, Z_d)$ *with probability distribution* $\mathbf{P}_Z$. *Let* $\mu_j = E[Z_j^1]$ *and* $\overline{Z}_j^n = 1/n \sum_{t=1}^n Z_j^t, j = 1, \ldots, d$ *denote the expectation and sample mean, respectively of component* Z_j.

Assume that $\forall \varepsilon > 0$, $j = 1, \ldots, d$, *there exist numbers* $\alpha_j, \gamma_j > 0$ *such that* $\mathbf{P}_Z[|\overline{Z}_j^n - \mu_j| > \varepsilon] \leq \alpha_j e^{-\gamma_j n}, \forall n \geq 1$.

Then, for any continuous function $F(z_1, \ldots, z_d)$, $\forall \varepsilon > 0$, $\exists \alpha, \gamma > 0$, *such that* $\mathbf{P}_Z[|F(\overline{Z}_1^n \ldots, \overline{Z}_d^n) - F(\mu_1, \ldots, \mu_d)| > \varepsilon] \leq \alpha e^{-\gamma n}, \forall n \geq 1$.

Proof: Let $A(n, \varepsilon) = \{|F(\overline{Z}_1^n, \ldots, \overline{Z}_d^n) - F(\mu_1, \ldots, \mu_d)| > \varepsilon\}$. Since F is continuous, the event $A(n, \varepsilon)$ implies the event

$$\{|\overline{Z}_j^n - \mu_j| > \zeta_j , \quad \text{for some } j = 1, \ldots, d\}$$

for suitable $\zeta_j = \zeta_j(\varepsilon) > 0$. Therefore,

$$\mathbf{P}_Z A(n, \varepsilon) \leq \sum_{j=1}^{d} \mathbf{P}_Z[|\overline{Z}_j^n - \mu_j| > \zeta_j] \leq \sum_{j=1}^{d} \alpha_j(\zeta_j) e^{-\gamma_j(\zeta_j) n} \leq \alpha e^{-\gamma n} ,$$

where $\gamma = \min_j \gamma_j(\zeta_j)$ and $\alpha = \sum_{j=1}^d \alpha_j(\zeta_j)$. This complete the proof. □

Proposition 4:

1. $\forall \varepsilon > 0$, $\exists \alpha^g$, $\gamma^g > 0$ *such that* $\mathbf{P}[E_n^g > \varepsilon] \leq \alpha^g e^{-\gamma^g n}$ $\forall n \geq 1$.
2. *For* $j = m, w, s$, $\forall x \in S$, $\forall \varepsilon > 0$, $\exists \alpha^j(x), \gamma^j(x) > 0$ *such that*

$$\mathbf{P}[E_n^j(x) > \varepsilon] \leq \alpha^j(x) e^{-\gamma^j(x) n} \quad \forall n \geq 1 .$$

Proof: (1) From Proposition 2 it follows that $0 \leq E_n^g = U_n^g - L_n^g = F(\bar{I}_n(0), \ldots, \bar{I}_n(s), \bar{T}_n(0), \ldots, \bar{T}_n(s), \overline{W}_n(0), \ldots, \overline{W}_n(s))$, where F is a continuous function, with $F(p(0), \ldots, p(s), m(0)p(0), \ldots, m(s)p(s), w(0)p(0), \ldots, w(s)p(s)) = 0$. Therefore, the assertion follows from Lemmata 2, 3.

The proof of part 2 is similar. □

References

Burnetas AN, Katehakis MN (1996) Finding optimal policies for markovian decision processes using simulation. Prob. Eng. Info. Sci. 10:525–537

Burnetas AN, Katehakis MN (1997) Optimal adaptive policies for markovian decision processes. Math. Oper. Res. 22(1):222–255

Dembo A, Zeitouni O (1993) Large deviations techniques and applications. Jones and Bartlett

Derman C (1970) Finite state Markovian decision processes. Academic Press

Federgruen A, Schweitzer P (1981) Nonstationary markov decision problems with converging parameters. J. Opt. Th. Appl. 34:207–241

Fishman GS (1978) Principles of discrete event simulation. Wiley

Fishman GS (1994) Markov chain sampling and the product estiamtor. Mgt. Sci. 42:1137–1145

Glasserman P, Liu T (1996) Rare-event simulation for multistage production-inventory systems. Mgt. Sci. 42(9):1292–1307

Hernández-Lerma O (1989) Adaptive Markov control processes. Springer-Verlag

Hordijk A (1974) Dynamic programming and Markov potential theory. Mathematisch Centrum, Amsterdam

Hordijk A, Iglehart DL, Schassberger R (1976) Discrete time methods for simulating continuous time markov chains. Adv. App. Prob. 8:772–788

Katehakis MN, Robbins H (1995) Sequential allocation involving normal populations. Proc. Natl. Acad. Sci. USA:8584–8585

Kleijnen JPC (1992) Simulation: A statistical perspective. Chichester

Van-Dijk N, Puterman ML (1988) Perturbation theory for Markov reward processes with applications to queueing systems. Adv. App. Prob. 20:79–98

Received: October 1996
Revised version received: May 1997