On Finding Optimal Policies for Markovian Decision Processes Using Simulation

Apostolos N. Burnetas
Case Western Reserve University

Michael N. Katehakis
Rutgers University

February 1995

Abstract

A simulation method is developed, to find an optimal policy for the expected average reward of a Markovian Decision Process. It is shown that the method is consistent, in the sense that it produces solutions arbitrarily close to the optimal. Various types of estimation errors are examined, and bounds are developed.

1. Introduction. Consider the problem of finding optimal policies for a finite state/action space Markovian Decision Process, when the criterion is the expected average reward. Various computational methods have been proposed for this problem, such as value iteration, policy improvement, linear programming etc..

In this paper we develop a simulation algorithm. The algorithm substitutes the value determination phase of the Policy Improvement method (i.e. the solution of typically large scale systems of linear equations) with simulation, and it is readily implementable in an environment of parallel processing. It is shown that it produces policies with expected average reward arbitrarily close to the optimal.

Specifically, the method is based on simulating independent cycles of the underlying Markov process for a policy under consideration. We develop estimates of the average and differential rewards, as well as of the test quantities that are used in the policy improvement step. However, since only estimates of these quantities are available, the estimation errors may affect the outcomes of the optimality testing and the improvement step. We solve this problem by obtaining bounds for the test quantities and using these instead of the estimates. These bounds are simultaneous, hold with probability one and converge to the true values as the number of samples increases.

In section 2 we describe the problem and outline the policy improvement method and the optimality equations. In section 3 we develop an iterative simulation algorithm which converges almost surely to at least an $\epsilon -$ optimal policy in finite number of steps. The convergence of the algorithm is assured given the existence of an estimation procedure which can be used in place of the policy evaluation phase to produce a set of bounds for the test quantities that satisfy almost sure convergence conditions. In section 4 we describe such

an estimation procedure which as we prove satisfies the necessary conditions for the correct behavior of the algorithm.

The derivation of probability one bounds is based on the following important observation. The estimation errors of the average and differential rewards associated with the policy under consideration satisfy the policy evaluation equations of a problem with the same transition mechanism and suitably modified reward structure. Therefore they are themselves expected average and differential rewards for the modified problem.

The simulation and estimation mechanism is based on the regenerative simulation method proposed in **?** and **?** for a problem without decisions. Bounds for the optimal expected average reward as well as the finite horizon and discounted rewards, when the transition matrix is subject to small perturbations are derived in **?**. **?** report computational comparisons among various methods using bound approximations and action elimination in the value iteration for the discounted problem. **?** develop an adaptive estimation mechanism for the problem of unknown transition and/or reward structure. For additional information on the related topic of adaptive control of Markov Processes the reader is referred to **?**, **?**, **?** and references therein.

2. The Model-Background. Consider a Markovian Decision Process described by $\{S, A(x), x \in S, P, R\}$, where S is the state space, $A(x)$ is the action set in state x, $x \in S$, $P = (p_{xy}(a))_{x,y \in S, a \in A(x)}$ is the transition mechanism, and $R = (r(x,a))_{x \in S, a \in A(x)}$ is the reward structure.

A policy π is generally defined as a random mechanism

$$P[A_t = a_t] = \pi_t(a_t | x_0, a_0, \ldots, x_{t-1}, a_{t-1}, x_t).$$

A policy π is *deterministic* if there exists a sequence of functions $\{f_t : S \to \bigcup_{x \in S} A(x), t = 0, 1, \ldots, \}$, such that $f_t(x) \in A(x), \forall x, t$ and

$$\pi_t(a_t | x_0, a_0, \ldots, x_{t-1}, a_{t-1}, x_t) = 1 \text{ or } 0$$

according to whether $a_t = f_t(x_t)$ or not. A policy is *stationary* if π_t, f_t do not depend on t. Let D denote the set of all policies and $\mathcal{D}$ the set of all *deterministic stationary* policies.

The *expected long term average reward* associated with policy π and initial state $X_0 = x$ is defined as $g(x, \pi) = \limsup_{N \to \infty} \frac{1}{N+1} \sum_{t=0}^{N} E(r(X_t, A_t) | x, \pi)$.

A policy π_0 is *optimal* with respect to the expected average reward criterion if $g(x, \pi_0) \geq g(x, \pi), \forall x \in S, \pi \in D$.

For each policy $f \in \mathcal{D}$ and initial state $X_0 = x_0 \in S$, the resulting stochastic process $\{X_t, t = 0, 1, \ldots\}$ is Markov with state space S, transition matrix $P(f) = (p_{xy}(f(x)))_{x,y \in S}$, and reward function $r(f) = (r(x, f(x)))_{x \in S}$.

Let $p_x(a), p_x(f)$ denote the probability row vectors $(p_{xy}(a))_{y \in S}$, $(p_{xy}(f(x)))_{y \in S}$, respectively. Also for any function $h(x)$ defined on S, let h denote the column vector $(h(x))_{x \in S}$, and $p_x(a)\, h$ the inner product $\sum_{y \in S} p_{xy}(a)\, h(y)$.

Assumption 2.1 We assume the following.

1. The state and action spaces $S, A(x),\ x \in S$ are finite sets.

2. $P(f)$ is irreducible for all $f \in \mathcal{D}$.

Under assumption 2.1 the expected average rewards associated with a policy do not depend on the initial state. For this case it is well known (e.g. **?**, **?**) that there exists an optimal deterministic stationary policy f, if and only if there exist bounded solutions to the following system of functional equations

$$g + h(x) = \max_{a \in A(x)} \{r(x,a) \ + \ p_x(a)\, h\},\ x \in S. \tag{2.1}$$

Optimal policies are defined by the maximizing actions in the above equations.

The Policy Improvement method starts with an arbitrary policy $f_0 \in \mathcal{D}$ at iteration 0. Iteration n proceeds as follows.

1. At the policy evaluation phase compute $g_{f_n}, h_{f_n}(x),\ x \in S$ as the unique solution to the following system of linear equations,

$$g_{f_n} + h_{f_n}(x) = r(x, f_n(x)) + p_x(f_n)\, h_{f_n}\ , x \in S. \tag{2.2}$$

   where, we use the normalization $h_{f_n}(0) = 0$, for a fixed but arbitrarily chosen state 0 in S.

2. At the policy improvement phase compute the *test quantities* $\phi_{f_n}(x,a),\ x \in S, a \in A(x)$

$$\begin{aligned} \phi_{f_n}(x,a) &= g_{f_n} + h_{f_n}(x) - (r(x,a) + p_x(a)\, h_{f_n}) \\ &= r(x, f_n(x)) - r(x,a) + (p_x(f_n) - p_x(a))h_{f_n}. \end{aligned} \tag{2.3}$$

3. The current policy f_n is optimal iff

$$\phi_{f_n}(x,a) \geq 0\,,\ \forall\ x \in S, a \in A(x). \tag{2.4}$$

4. If (2.4) is violated at some states, a new policy f_{n+1} , not necessarily uniquely defined, for iteration $n+1$ is constructed by

$$f_{n+1}(x) = \arg \min_{a \in A(x)} \phi_{f_n}(x,a). \tag{2.5}$$

3. The Simulation Algorithm. Recall that termination – computation of improved policies at every step of the Policy Improvement procedure is determined by the optimality conditions (2.4) : $\phi_f(x,a) \geq 0, \forall\ x \in S, a \in A(x)$.

Instead of solving (2.2), we perform simulation to obtain estimates $\hat{g}_f, \hat{h}_f(x), \hat{\phi}_f(x,a)$ of $g_f, h_f(x), \phi_f(x,a)$, and use them for the optimality testing in (2.4). The estimation errors however may cause an incorrect conclusion of an optimality testing of the form $\hat{\phi}_f(x,a) \geq 0$, and of the corresponding improvement step. This problem can be overcome if we modify

the optimality/improvement steps, using suitably defined bounds for $\phi_f(x,a)$ instead of the estimates $\hat{\phi}_f(x,a)$.

In this section we postulate the existence of an estimation procedure which, for any policy $f \in \mathcal{D}$, computes estimates $\hat{g}_f, \hat{h}_f(x), \hat{\phi}_f(x,a)$ of $g_f, h_f(x), \phi_f(x,a)$, and random variables $L_t^{\phi,f}(x,a), U_t^{\phi,f}(x,a)$ with the following properties

($\mathbf{P}_1$) $L_t^{\phi,f}(x,a) \leq \phi_f(x,a) \leq U_t^{\phi,f}(x,a), \quad \forall x \in S, a \in A(x),$

($\mathbf{P}_2$) $L_t^{\phi,f}(x,a), U_t^{\phi,f}(x,a) \to \phi_f(x,a)$ with probability 1 as $t \to \infty$, where t denotes the number of simulation transitions.

An estimation procedure that satisfies $\mathrm{P}_1, \mathrm{P}_2$ is presented in the next section.

We show that, under this assumption, there exists an $\epsilon-$optimal simulation algorithm, i.e. for any $\epsilon > 0$ there exists a simulation algorithm $A(\epsilon)$ which after an almost surely finite number of transitions produces a policy f such that $g_{f^*} - g_f < \epsilon$, where f^* is an optimal policy.

Let

$$M_u(f,t) = \min_{x \in S, a \neq f(x)} U_t^{\phi,f}(x,a), \tag{3.1}$$

$$M_l(f,t) = \min_{x \in S, a \neq f(x)} L_t^{\phi,f}(x,a). \tag{3.2}$$

For any $\epsilon > 0$, algorithm $A(\epsilon)$ is defined as follows.

Iteration 0 At iteration $n = 0$ select an arbitrary policy $f_0 \in \mathcal{D}$.

Iteration n At iteration n, with policy $f_n \in \mathcal{D}$,

1. Apply the simulation procedure for the minimum number t of transitions required, so that one of the following conditions is satisfied.

(C-a) $M_u(f_n,t) < 0$,

(C-b) $M_l(f_n,t) > 0$, or

(C-c) $M_l(f_n,t) > -\epsilon$.

2. If condition C-a is satisfied, then

Construct policy f_{n+1} as follows

$$f_{n+1}(x) = \begin{cases} \arg\min_{a \in A(x)} U_t^{\phi,f_n}(x,a) & , \text{ if } \min_{a \in A(x)} U_t^{\phi,f_n}(x,a) < 0 \\ f_n(x) & , \text{ otherwise.} \end{cases} . \tag{3.3}$$

Return to Iteration $n+1$ with policy f_{n+1}.

If condition C-b is satisfied, stop and return f_n as the unique optimal policy.

If condition C-c is satisfied, stop and return f_n as an ϵ–optimal policy.

Let $f_1, f_2 \in \mathcal{D},\ d^g = g_{f_2} - g_{f_1},$ and $d^h(x) = h_{f_2}(x) - h_{f_1}(x),\ x \in S.$

In order to analyze the convergence of algorithm $A(\epsilon)$ we need the following lemma.

Lemma 3.1

1. *For any two policies $f_1, f_2 \in \mathcal{D}$ the quantities $d^g, d^h(x), \phi_{f_1}(x, f_2(x)),\ x \in S$ satisfy the following relations*

$$d^g + d^h(x) = -\phi_{f_1}(x, f_2(x)) + p_x(f_2)\, d^h, \quad x \in S. \tag{3.4}$$

2. *If for some policy $f \in \mathcal{D}$ there exists $\zeta > 0$, such that $\phi_f(x, a) > -\zeta, \forall\ x \in S, a \in A(x)$, then*

$$g_{f^*} - g_f < \zeta. \tag{3.5}$$

Proof. To prove part 1, consider the policy evaluation equations for policy f_2

$$g_{f_2} + h_{f_2}(x) = r(x, f_2(x)) + p_x(f_2)h_{f_2},\ x \in S. \tag{3.6}$$

Using the definitions of $d^g, d^h(x)$, the above relation can be rearranged as follows

$$d^g + d^h(x) \quad = \quad -(g_{f_1} + h_{f_1}(x) - r(x, f_2(x)) - p_x(f_2)\, h_{f_1}) + p_x(f_2)\, d^h. \tag{3.7}$$

The quantity inside the parentheses on the right hand side of (3.7) is equal to $\phi_{f_1}(x, f_2(x))$ by definition. Thus (3.4) follows.

We next prove part 2. A direct consequence of part 1 is that the difference d^g is equal to the expected average reward of a Markov Reward Process with transition matrix $P(f_2)$ and rewards are equal to the opposite of the test quantities $\phi_{f_1}(x, f_2(x))$. Therefore

$$d^g = -\sum_{x \in S} \pi_{f_2}(x)\phi_{f_1}(x, f_2(x))\ ,$$

where π_{f_2} is the steady state probability vector of $P(f_2)$. Thus if $\phi_{f_1}(x, a) > -\zeta,\ x \in S, a \in A(x)$, then $d^g < \zeta$. (3.5) follows letting $f_1 = f, f_2 = f^*$. □

Theorem 3.2 *For any $\epsilon > 0$, under the assumption of existence of an estimation scheme satisfying properties P_1 and P_2, algorithm $A(\epsilon)$ stops with probability one after a finite total number of transitions, with a policy f such that $g_{f^*} - g_f < \epsilon$, where f^* is an optimal policy.*

Proof. Since the total number of deterministic stationary policies is finite, it suffices to show that , for any n, iteration n of algorithm $A(\epsilon)$ will stop with probability one after a finite number of simulation transitions and either it will correctly conclude that f_n is $\epsilon -$ optimal, i.e. $g^*_f - g_{f_n} < \epsilon$, or it will produce a new policy f_{n+1}, to be used in iteration $n + 1$, such that $g_{f_{n+1}} - g_{f_n} > 0$.

In order to prove this claim, consider iteration n and distinguish the following cases.

Case 1. Assume that at iteration n policy f_n satisfies $\min_{x\in S, a\in A(x)} \phi_{f_n}(x,a) \geq 0$, i.e it is optimal.

Since $U_t^{\phi,f_n} \geq \phi_{f_n}(x,a)$, the algorithm will always find $U_t^{\phi,f_n}(x,a) \geq 0,\ \forall x \in S, a \in A(x)$, therefore iteration n will not stop under condition C-a.

In addition, since $L_t^{\phi,f_n}(x,a) \to \phi_{f_n}(x,a) \geq 0$ w.p. 1, as $t \to \infty$, it follows that for all $x \in S, a \in A(x), a \neq f_n(x)$ there exists, w.p. 1, a finite number $t(x,a)$ such that $L_t^{\phi,f_n}(x,a) > -\epsilon,\ \forall\, t > t(x,a)$. Thus $M_l(f_n,t) > -\epsilon\ \forall t > t_0$, where $t_0 = \max t(x,a)$.

The last inequality means that there exists, w.p. 1, a finite number t_0, such that after at most t_0 transitions within iteration n, one of the conditions C-b and C-c will be satisfied, and iteration n will terminate either with the conclusion that f_n is optimal, or that it is $\epsilon-$ optimal, both of which conclusions are correct.

Case 2. Assume that at iteration n policy f_n satisfies $-\epsilon \leq \min_{x\in S, a\in A(x)} \phi_{f_n}(x,a) < 0$, i.e., from Lemma 3.1, f_n is $\epsilon-$ optimal.

In this case the algorithm will always find that $L_t^{\phi,f_n}(x,a) < 0$ for at least one pair (x,a), thus iteration n will not stop under condition C-b.

Because $L_t^{\phi,f_n} x, a, U_t^{\phi,f_n}(x,a) \to \phi_{f_n}(x,a)$ w.p. 1, it follows with the same reasoning as in case 1, that, w.p. 1, after a finite number of transitions within iteration n, either condition C-a or condition C-c will be satisfied and the iteration will stop. If it stops under condition C-c, the correct claim that f_n is $\epsilon-$optimal will be made. If it stops under C-a, then a new policy f_{n+1} will be derived. By definition of f_{n+1} we have that $U_t^{\phi,f_n}(x, f_{n+1}(x)) < 0,\ \forall x$ such that $f_{n+1}(x) \neq f_n(x)$, thus,

$$\phi_{f_n}(x, f_{n+1}(x)) \leq 0,\ \forall x \in S,$$

with strict inequality for at least one state x. Therefore from Lemma (3.1) it follows that $g_{f_{n+1}} - g_{f_n} > 0$.

Case 3. Assume that at iteration n policy f_n satisfies $\min_{x\in S, a\in A(x)} \phi_{f_n}(x,a) < -\epsilon$.

In this case the algorithm will always find that $L_t^{\phi,f_n}(x,a) < -\epsilon\ \forall t$, for at least one pair (x,a), thus iteration n will not stop under conditions C-b or C-c.

Since $U_t^{\phi,f_n}(x,a) \to \phi_{f_n}(x,a) < -\epsilon$ w.p. 1, for at least one pair (x,a), it follows the same way as in case 1, that after a finite number of transitions a new policy f_{n+1} will be produced, which is strictly improved with respect to f_n.

Therefore the claim has been proved for all cases, and the proof of the theorem is complete. □

4. The Estimation Procedure. In this section we develop an estimation procedure which satisfies properties P_1 and P_2 described in section 3, and therefore assures the correct behavior of algorithm $A(\epsilon)$. The procedure works as follows.

Select a fixed state 0 which from now on will be used as the state of reference. A cycle is defined as the number of transitions between two successive returns to state 0. At iteration n of the algorithm $A(\epsilon)$, with policy f_n under consideration, the Markov Reward Process with transition matrix $P(f_n)$ and reward vector $r(f_n) = r(x, f_n(x)),\ x \in S$ is simulated for a number of cycles. During the simulation estimates of $g_{f_n}, h_{f_n}(x)$ are calculated, using an estimation scheme which will be described in detail. Using these estimates and based on a number of intermediate results, we show the existence of lower and upper bounds $L_t^{\phi,f_n}(x,a), U_t^{\phi,f_n}(x,a)$ for the test quantities $\phi_{f_n}(x,a)$, that satisfy $\mathrm{P}_1, \mathrm{P}_2$.

We first define a set of random variables which will be used in the development of the estimation method.

Let $f = f_n$ be the policy used in iteration n of the algorithm. Index t denotes the step number (simulation transition) within any iteration n. Index j denotes the j^{th} cycle within any iteration n. A cycle is considered as a sample of our estimation scheme, therefore j also represents the j^{th} sample within a single iteration. Both j and t are reset to zero at the beginning of each iteration n.

Let P_x^f, E_x^f denote the probability distribution and the expectation respectively, with respect to transition matrix $P(f)$ and initial state x.

Let $\beta_0^f = 0$, and

$$\beta_j^f = \min\{t \geq \beta_{j-1}^f + 1, X_t = 0, X_i \neq 0, i = \beta_{j-1}^f + 1, \ldots, t-1\}, \quad j = 1, 2, \ldots \tag{4.1}$$

denote the successive return epochs to state 0, when policy $f \in \mathcal{D}$ is used.

Also let

$$A_j^f(x) = \min\{t : \beta_j^f \leq t \leq \beta_{j+1}^f - 1, X_t = x\}. \tag{4.2}$$

represent the epoch of first visit to state x between the j^{th} and the $(j+1)^{st}$ recurrence to state 0, under policy f, assuming that $\min \emptyset = +\infty$.

Define

$$I_j^f(x) = 1\{A_j^f(x) < \infty\} = \begin{cases} 1 & , \text{ if state } x \text{ is visited during the } j^{th} \text{ cycle} \\ 0 & , \text{ otherwise} \end{cases}, \tag{4.3}$$

$$T_j^f(x) = I_j^f(x)(\beta_{j+1}^f - A_j^f(x)), \tag{4.4}$$

$$W_j^f(x) = I_j^f(x) \sum_{t=A_j^f(x)}^{\beta_{j+1}^f - 1} r(X_t, f(X_t)). \tag{4.5}$$

The random variables $T_j^f(x), W_j^f(x)$ represent the number of transitions and the reward obtained in the time interval between the first visit to state x and the next recurrence to state 0, during the j^{th} cycle. If x is not visited in the j^{th} cycle, they are set equal to 0 [1].

[1] It is easy to see that $I_j^f(0) = 1, A_j^f(0) = \beta_k^f, T_j^f(0) = \beta_{j+1}^f - \beta_j^f$.

Remark 4.1 As a direct consequence of the fact that every state of a positive recurrent Markov chain is a regeneration point of the process, we observe that the random vectors

$$V_j = \{\beta_j^f, I_j^f(x), A_j^f(x), T_j^f(x), W_j^f(x), \}\ \ x \in S, j = 1, 2, \ldots$$

are independent and identically distributed. This property will be used to prove the consistency of our estimation procedure.

Lemma 4.2 *The quantities $g_f, h_f(x),\ x \in S$ are given by*

$$g_f = \frac{E^f[W_j^f(0)]}{E^f[T_j^f(0)]} \tag{4.6}$$

and

$$h_f(x) = \frac{E^f[W_j^f(x)] - g_f E^f[T_j^f(x)]}{P^f[I_j^f(x) = 1]}. \tag{4.7}$$

Proof. From **?**, p.66 (see also **?**, p.126)) we obtain the following interpretation of $g_f, h_f(x)$, as conditional expectations given a policy f and an initial state x

$$g_f = \frac{E_0^f[\sum_{t=0}^{\beta_1^f - 1} r(X_t, f(X_t))]}{E_0^f[\beta_1^f]}, \tag{4.8}$$

and

$$\begin{aligned} h_f(x) &= E_x^f[\sum_{t=0}^{\beta_1^f - 1} (r(X_t, f(X_t)) - g_f)] \\ &= E_x^f[\sum_{t=0}^{\beta_1^f - 1} r(X_t, f(X_t))] \ - \ g_f E_x^f[\beta_1^f]. \end{aligned} \tag{4.9}$$

From the definition of $W_j^f(x)$ we see that $\forall\ x \in S$

$$\begin{aligned} E^f[W_j^f(x)] &= E^f[E^f[W_j^f(x)/I_j^f(x)]] \\ &= P^f[I_j^f(x) = 1]\, E^f[W_j^f(x)/I_j^f(x) = 1]. \end{aligned} \tag{4.10}$$

But

$$\begin{aligned} E^f[W_j^f(x)/I_j^f(x) = 1] &= E^f[W_1^f(x)/I_1^f(x) = 1] \\ &= E_x^f[\sum_{t=0}^{\beta_1^f - 1} r(X_t, f(X_t))], \end{aligned} \tag{4.11}$$

where the first equality follows from Remark (4.1) and the second from the Markov property.

In a similar manner we can show that

$$E^f[T_j^f(x)] = P^f[I_j^f(x) = 1]\, E^f[\beta_1^f]. \tag{4.12}$$

Since all states are positive recurrent, $P^f[I_j^f(x) = 1] > 0,\ x \in S$.

Relations (4.6) and (4.7) follow from (4.8)–(4.12). □

Assume that k samples $V_1, \ldots, V_k$ of the random vector V, defined in Remark (4.1), have been obtained, after simulating the process for k cycles using policy f. Then we form the following estimates for the expected average and differential rewards and the test quantities associated with f

$$\hat{g}_f = \frac{\overline{W}_k^f(0)}{\overline{T}_k^f(0)}, \tag{4.13}$$

$$\hat{h}_f(x) = \frac{\overline{W}_k^f(x) - \hat{g}_f \overline{T}_k^f(x)}{\overline{I}_k^f(x)},\ x \in S, \tag{4.14}$$

$$\hat{\phi}_f(x,a) = r(x,f(x)) - r(x,a) + (p_x(f) - p_x(a))\, \hat{h}_f, \tag{4.15}$$

where $\overline{I}_k^f = \sum_{j=1}^k I_j x/k,\ \overline{T}_k^f = \sum_{j=1}^k T_j^f(x)/k,\ \overline{W}_k^f \sum_{j=1}^k W_j^f(x)/k$. If $\overline{I}_k^f(x) = 0$, we set $\hat{h}_f(x)$ arbitrarily. Note that $\hat{h}_f(0) = 0$, which is consistent with the adopted normalization.

For notational simplicity we do not specifically indicate the dependence of the estimates $\hat{g}_f, \hat{h}_f, \hat{\phi}_f$ on the sample size k. This will also be true for all the quantities defined as functions of the estimates throughout this section.

Lemma 4.3 *$\hat{g}_f$ and $\hat{h}_f(x), \hat{\phi}_f(x,a),\ x \in S, a \in A(x)$ are strongly consistent estimates of g_f and $h_f(x), \phi_f(x,a),\ x \in S, a \in A(x)$ respectively.*

Proof. From Remark (4.1) and the strong law of large numbers we get that $\overline{W}_k^f(x) \to E^f[W_j^f(x)], \overline{T}_k^f(x) \to E^f[T_j^f(x)], \overline{I}_k^f(x) \to E^f[I_j x] = P^f[I_k x = 1]$, w.p. 1 as $n \to \infty$. Note that all the involved expectations are finite, because of the bounded rewards and the irreducibility assumption. The lemma now follows from lemma (4.2). □

In order to develop bounds $L_f(x,a), U_f(x,a)$ for $\phi_f(x,a)$ we need two intermediate results which are presented in Lemma 4.4 and Lemma 4.5 below. Although the results of Lemma 4.4 are contained in **?**, the proof is presented here for completeness.

Let

$$m^f(x) = E_x^f[\beta_1^f], \tag{4.16}$$

denote the expected first passage time from state x to state 0 under policy f. The quantities $m^f(x),\ x \in S$ are the unique solution to the following system of equations

$$m^f(x) = 1 + \sum_{y \neq 0} p_{xy}(f(x))\, m^f(y),\ x \in S. \tag{4.17}$$

Also let

$$\hat{m}^f(x) = \frac{\overline{T}_k^f(x)}{\overline{I}_k^f(x)},\ x \in S, \tag{4.18}$$

and

$$\begin{aligned} \eta^{m,f}(x) &= \hat{m}^f(x) - (1 + \sum_{y \neq 0} p_{xy}(f(x))\hat{m}^f(y)), & (4.19) \\ \eta_{min}^{m,f} &= \min_{x \in S} \eta^{m,f}(x). & (4.20) \end{aligned}$$

$\hat{m}^f(x)$ is a strongly consistent estimate of $m^f(x)$. $\eta^{m,f}(x)$ represents the difference between the left and right hand side of (4.17), when the estimates $\hat{m}^f(x)$ are used instead of the true values $m^f(x)$.

Lemma 4.4 *The quantities $m^f(x),\ x \in S$ are bounded as follows*

$$m^f(x) \leq U^{m,f}(x), \tag{4.21}$$

where

$$\begin{aligned} U^{m,f}(x) &= \hat{m}^f(x)(1 - \frac{\rho^{m,f}}{1 + \rho^{m,f}}), & (4.22) \\ \rho^{m,f} &= \max\{\eta_{min}^{m,f}, -1\}\ . & (4.23) \end{aligned}$$

Proof. Let $\delta^{m,f}(x) = \hat{m}^f(x) - m^f(x)$ denote the estimation error of $\hat{m}^f(x)$. Substituting $\hat{m}^f(x)$ and $\hat{m}^f(y)$ into (4.19) we get

$$\delta^{m,f}(x) = \eta^{m,f}(x) + \sum_{y \neq 0} p_{xy}(f(x))\delta^{m,f}(y). \tag{4.24}$$

Thus $\delta^{m,f}(x)$ represents the expected first passage reward from state x to 0 under reward structure $\eta^{m,f}(x)$, i.e.

$$\delta^{m,f}(x) = E_x^f[\sum_{t=0}^{\beta_1^f - 1} \eta^{m,f}(X_t)]. \tag{4.25}$$

Therefore the following inequality is immediate

$$\delta^{m,f}(x) \geq \eta_{min}^{m,f} m^f(x).$$

On the other hand since by definition $\hat{m}^f(x) > 0,\ x \in S$, it follows that

$$\delta^{m,f}(x) > -m^f(x).$$

Combining the last two inequalities we obtain

$$\delta^{m,f}(x)/m^f(x) \geq \rho^{m,f}. \tag{4.26}$$

We now observe that

$$\begin{aligned}\frac{\delta^{m,f}(x)}{\hat{m}^f(x)} &= \frac{\delta^{m,f}(x)}{m^f(x)+\delta^{m,f}(x)} \\ &= \frac{\delta^{m,f}(x)/m^f(x)}{1+\delta^{m,f}(x)/m^f(x)} \end{aligned} \tag{4.27}$$

Since the function $f(x) = x/(1+x)$ is increasing in x for $x > -1$, it follows from (4.26) and (4.27) that

$$\delta^{m,f}(x) \geq \frac{\rho^{m,f}\hat{m}^f(x)}{1+\rho^{m,f}},$$

from which (4.21) is immediate. □

Let

$$\delta^{g,f} = \hat{g}_f - g_f\,, \tag{4.28}$$

$$\delta^{h,f}(x) = \hat{h}_f(x) - h_f(x)\,, \tag{4.29}$$

represent the estimation errors of $\hat{g}_f, \hat{h}_f(x)$, and

$$\eta^{g,f}(x) = \hat{g}_f + \hat{h}_f(x) - (r(x,f(x)) + p_x(f)\,\hat{h}_f)\,. \tag{4.30}$$

the deviation of the policy evaluation equations (2.2) for policy f, when the estimates for g, h are used instead of the true values. Also let

$$\eta^{g,f}_{min} = \min_{x\in S} \eta^{g,f}(x),$$
$$\eta^{g,f}_{max} = \max_{x\in S} \eta^{g,f}(x).$$

Lemma 4.5

1. *The error $\delta^{g,f}$ is bounded as follows*

$$\eta^{g,f}_{min} \leq \delta^{g,f} \leq \eta^{g,f}_{max}. \tag{4.31}$$

2. *The error $\delta^{h,f}(x)$ is bounded as follows*

$$|\delta^{h,f}(x)| \leq U^{h,f}(x), \quad \forall x \in S, \tag{4.32}$$

where

$$\begin{aligned} U^{h,f}(x) &= (\eta^{g,f}_{max} - \eta^{g,f}_{min})U^{m,f}(x),\ x \neq 0, \\ U^{h,f}(0) &= 0. \end{aligned} \tag{4.33}$$

3. $U^{h,f}(x) \to 0$ *w.p. 1, as* $k \to \infty$, $\forall x \in S$.

Proof. The assertion is immediate for $x = 0$, because $h_f(0) = \hat{h}_f(0) = 0$. Let $x \neq 0$. From (4.28), (4.29), (4.30) we get

$$\begin{aligned} \eta^{g,f}(x) &= (g_f + \delta^{g,f}) + (h_f(x) + \delta^{h,f}(x)) - \ (r(x, f(x) + p_x(f)\,(h_f + \delta^{h,f})\,) \\ &= \delta^{g,f} + \delta^{h,f}(x) - p_x(f)\,\delta^{h,f}, \end{aligned}$$

hence,

$$\delta^{g,f} + \delta^{h,f}(x) = \eta^{g,f}(x) + p_x(f)\,\delta^{h,f}\ ,\ x \in S. \tag{4.34}$$

Thus the estimation errors satisfy a set of policy evaluation equations, with transition matrix $P(f)$ and reward at state $x \in S$ equal to $\eta^{g,f}(x)$. Let $\pi_f(x),\ x \in S$ denote the stationary probability vector of $P(f)$. Then

$$\delta^{g,f} = \sum_{x \in S} \pi_f(x) \eta^{g,f}(x), \tag{4.35}$$

therefore

$$\eta^{g,f}_{min} \leq \delta^{g,f} \leq \eta^{g,f}_{max}. \tag{4.36}$$

This proves (4.31).

In order to prove part 2 we observe that the quantities $\delta^{h,f}(x),\ x \in S$ play the role of the differential rewards for the modified process, and thus, according to (4.9),

$$\delta^{h,f}(x) = E^f_x[\sum_{t=0}^{\beta^f_1 - 1} (\eta^{g,f}(x) - \delta^{g,f})]. \tag{4.37}$$

Using (4.37) and (4.31) we derive the following bounds for $\delta^{h,f}(x)$.

$$(\eta^{g,f}_{min} - \eta^{g,f}_{max}) m^f(x) \leq \delta^{h,f}(x) \leq (\eta^{g,f}_{max} - \eta^{g,f}_{min}) m^f(x).$$

Now (4.32) can be shown by applying the results of Lemma 4.4 for $m^f(x)$ and observing that $\eta^{g,f}_{min} - \eta^{g,f}_{max} \leq 0$.

For part 3 we have that $\eta^{g,f}(x) \to 0,\ \forall x \in S$, w.p.1, as the sample size increases, which follows from Lemma 4.3. Therefore $\eta^{g,f}_{max} - \eta^{g,f}_{min} \to 0$, w.p. 1, as $k \to \infty$. □

We can now prove the main result of this section.

Let

$$L^{\phi,f}(x,a) = \hat{\phi}_f(x,a) - \rho^{\phi,f}(x,a), \tag{4.38}$$

$$U^{\phi,f}(x,a) = \hat{\phi}_f(x,a) + \rho^{\phi,f}(x,a), \tag{4.39}$$

where

$$\rho^{\phi,f}(x,a) = \sum_{y \in S} |p_{xy}(f(x)) - p_{xy}(a)| U^{h,f}(y). \tag{4.40}$$

Theorem 4.6 *The quantities $L^{\phi,f}(x,a), U^{\phi,f}(x,a)$ defined in (4.38), (4.39) satisfy properties P_1, P_2 in section 3, i.e.*

$$L^{\phi,f}(x,a) \leq \phi_f(x,a) \leq U^{\phi,f}(x,a), \quad \forall x \in S, a \in A(x), \tag{4.41}$$

and

$$L^{\phi,f}(x,a), U^{\phi,f}(x,a) \to \phi_f(x,a), \text{w.p. 1, as } k \to \infty. \tag{4.42}$$

Proof. Let $\delta^{\phi,f}(x,a) = \hat{\phi}_f(x,a) - \phi_f(x,a)$ denote the estimation error of $\hat{\phi}_f(x,a)$. Substituting $\phi_f(x,a), \hat{\phi}_f(x,a)$ from (2.3) and (4.15) we find

$$\delta^{\phi,f}(x,a) = (p_x(f) - p_x(a))\, \delta^{h,f}, \tag{4.43}$$

therefore

$$\begin{aligned} |\delta^{\phi,f}(x,a)| &\leq \sum_{y \in S} |p_{xy}(f(x) - p_{xy}(a)|\, |\delta^{h,f}(y)| \\ &\leq \sum_{y \in S} |p_{xy}(f(x) - p_{xy}(a)|\, U^{h,f}(y)|, \\ &= \rho^{\phi,f}(x,a), \end{aligned}$$

where the second inequality follows from Lemma 4.5.2. This proves (4.41).

In order to prove (4.42), we note that $\rho^{\phi,f}(x,a) \to 0$ w.p.1 as $k \to \infty$, $\forall x \in S, a \in A(x)$, which follows from Lemma 4.5.3. In addition, $hfi(x,a) \to \phi_f(x,a)$ w.p.1, because $\hat{\phi}_f$ is strongly consistent. Thus, $L^{\phi,f}(x,a), U^{\phi,f}(x,a) \to \phi_f(x,a)$, w.p.1 as $k \to \infty$. □

Remark 4.7 With respect to computational implementation of the algorithm, parallel simulation techniques can be readily applied. Since under any policy f successive cycles are independent, the estimation procedure can be distributed to different processors of a massively parallel system. Each one of a group of "simulator" processors performs simulation of recurrence cycles and computes one sample V_j from every cycle. This sample is then sent to a "data collecting" processor, which is responsible for combining the samples into the estimates, computing the bounds, testing for optimality and performing the policy improvement step. Each time a new iteration of algorithm $A(\epsilon)$ is started, all simulating processors are restarted with the new policy. Under this parallelization scheme, the simulator processors can work asynchronously, and the only synchronization is performed when a policy changes.

References

[1] R. Agrawal, D. Teneketzis, and V. Anantharam. Asymtotically efficient adaptive allocation schemes for controlled Markov chains : Finite parameter space. *IEEE-AC*, 12:12–99, 1989.

[2] M. A. Crane and D. L. Iglehart. Simulating stable stochastic systems, i: General multi–server queues. *Journal of the ACM*, 21:103–113, 1974.

[3] M. A. Crane and D. L. Iglehart. Simulating stable stochastic systems, ii: Markov chains. *Journal of the ACM*, 21:114–123, 1974.

[4] A. Federgruen and P. Schweitzer. Nonstationary Markov decision problems with converging parameters. *J. Optim. Theory Appl.*, 34:207–241, 1981.

[5] O. Hernández-Lerma. *Adaptive Markov Control Processes*. Springer–Verlag, 1989.

[6] P. R. Kumar. A survey of some results in stochastic adaptive control. *SIAM J. Control and Optimization*, 23:329–380, 1985.

[7] L.C. Thomas, R. Harley, and A.C. Lavercombe. Computational comparisons of value iteration algorithms for discounted Markov decision processes. *Oper. Res. Letters*, 2:72–76, 1983.

[8] N.M. Van-Dijk and M. L. Puterman. Perturbation theory for Markov reward processes with applications to queueing systems. 20:79–98, 1988.