Proceedings of the 17th World Congress
The International Federation of Automatic Control
Seoul, Korea, July 6-11, 2008

Stochastic Iterative Approximation for Parallel Rollout and Policy Switching

Hyeong Soo Chang *

* *Department of Computer Science and Engineering, Sogang University, Seoul, Korea (Tel: 82-2-705-8925; e-mail: hschang@sogang.ac.kr).*

Abstract: This paper considers stochastic iterative computation methods for approximately computing parallel rollout and policy switching policies, in the context of improving all available heuristic policies, for solving Markov decision processes and analyzes the convergence of the computation methods.

1. INTRODUCTION

Markov decision process (MDP) models (see, e.g., [4] [1] for substantial discussions) are widely used for modeling sequential decision-making problems that arise in engineering, economics, computer science, and the social sciences, etc. Consider an MDP $M = (X, A, P, R)$ with a finite state set X, a finite action set A, a bounded cost function $C : X \times A \to \mathbb{R}$, and a transition function P that maps $\{(x, a)|x \in X, a \in A\}$ to the set of probability distributions over X. We denote the probability of making a transition to state $y \in X$ when taking action $a \in A$ in state $x \in X$ by P^a_{xy}. For simplicity, we assume that every action is admissible in every state.

Let Π be the set of all stationary policies $\pi : X \to A$. Define *the value of a policy* $\pi \in \Pi$ with an initial state $x \in X$:

$$V^\pi(x) = E\left[\sum_{t=0}^{\infty} \gamma^t C(x_t, \pi(x_t)) \middle| x_0 = x\right], x \in X,$$

where x_t is a random variable denoting state at time t and $\gamma \in (0, 1)$ is a discount factor.

Suppose that we have a nonempty set $\Delta \subseteq \Pi$ of heuristic policies for the control of the MDP M. The following two multi-policy improvement methods (called parallel rollout and policy switching [2]) provide a policy whose performance is no worse than any policy in Δ, respectively. We formally define a parallel rollout policy $\pi_{\rm pr}$ as

$$\pi_{\rm pr}(x) \in \arg\min_{a \in A}\left\{C(x, a) + \gamma \sum_{y \in X} P^a_{xy} \min_{\pi \in \Delta} V^\pi(y)\right\} \quad (1)$$

for $x \in X$ and define a policy switching policy $\pi_{\rm ps}$ as

$$\pi_{\rm ps}(x) \in \arg\min_{\pi \in \Delta}(V^\pi(x))(x), x \in X. \quad (2)$$

In words, at each state $x \in X$, the policy switching policy prescribes the action $\pi_{\rm ps}(x)$ prescribed by the policy that achieves $\min_{\pi \in \Delta}(V^\pi(x))$. It has been shown [2] that

$$V^{\pi_{\rm pr}}(x) \leq \min_{\pi \in \Delta} V^\pi(x), x \in X$$

and similarly,

$$V^{\pi_{\rm ps}}(x) \leq \min_{\pi \in \Delta} V^\pi(x), x \in X.$$

A stochastic iterative variant of the well-known policy iteration (PI) algorithm [4] for solving MDPs and its convergence have been studied by Tsitsiklis based on a *single* policy improvement in an "optimistic" way [6]. The problem context is different here from that of Tsitsiklis. Tsitsiklis' method is for computing an optimal policy in the entire set Π. It is often true that for a given problem, we already have a set of some heuristic policies available (for on-line control in some cases). For example, for the multiclass-scheduling problem with stochastically arriving prioritized tasks with deadlines, the "earliest-deadline-first" and "static-priority" heuristics are available candidate policies in hand for the scheduling decision. It may even be the case that our heuristic policies are such that each policy is near-optimal over some part of the state space. In this case, the decision maker may well wish to combine those policies to develop a policy that somehow improves all of the heuristic policies. Parallel rollout and policy switching have been studied in this context [2] [3] and this paper considers stochastic iterative computation methods, based on the idea of Tsitsiklis, for approximately computing parallel rollout and policy switching policies and analyzes the convergence of the computation methods.

Even though the optimistic PI considered in [6] converges to an optimal policy π^*, due to the very optimistic computation, a policy $\phi_t \in \Pi$ generated at iteration t in the optimistic PI does not necessarily improve ϕ_{t-1} unlike the monotonicity of the policies in the original PI. In other words, it is not necessarily true that $V^{\phi_t}(x) \leq V^{\phi_{t-1}}(x), \forall x \in X$, which means that after t-iterations, it cannot be guaranteed for ϕ_t to improve the policies generated at the previous iterations. Furthermore, it appears nontrivial to analyze the relative error bound between $V^{\phi_t}(x)$ and $V^{\pi^*}(x)$ for $x \in X$ and no analysis has been provided for such a finite-time bound of the optimistic PI. This paper provides a finite-time error bound between the value of the approximate parallel rollout policy (policy switching policy) and $\min_{\pi \in \Delta} V^\pi(x), x \in X$.

2. STOCHASTIC ITERATIVE COMPUTATION

Let $B(X)$ be the space of real-valued bounded functions on X. We define an operator $T : B(X) \to B(X)$ as

$$T(\Phi)(x) = \min_{a \in A} \left\{ C(x,a) + \gamma \sum_{y \in X} P_{xy}^{a} \Phi(y) \right\}$$

for $\Phi \in B(X), x \in X$ and similarly, an operator $T_\pi : B(X) \to B(X)$ for $\pi \in \Pi$ as

$$T_\pi(\Phi)(x) = C(x, \pi(x)) + \gamma \sum_{y \in X} P_{xy}^{\pi(x)} \Phi(y), x \in X$$

for $\Phi \in B(X)$. It is well known (see, e.g., [4]) that for each policy $\pi \in \Pi$, there exists a corresponding unique $\Phi \in B(X)$ such that for $x \in X$,

$$T_\pi(\Phi)(x) = \Phi(x) \text{ and } \Phi(x) = V^\pi(x).$$

2.1 Parallel Rollout

Monte Carlo policy evaluation Based on the parallel-rollout multi-policy improvement method in (1), we consider the following optimistic variant of it: at each iteration $t \geq 0$, we have available value functions $V_t^\pi, \pi \in \Delta$, defined over X and a value function J_t defined by

$$J_t(x) = \min_{\pi \in \Delta} V_t^\pi(x), x \in X.$$

Let $\mu_t^{\text{pr}} \in \Pi$ be a corresponding greedy policy such that

$$T_{\mu_t^{\text{pr}}}(J_t)(x) = T(J_t)(x), x \in X. \quad (3)$$

For each policy $\pi \in \Delta$ starting with each state $x \in X$, we generate a corresponding sample path over infinite horizon starting with state x and observe its cumulative discounted cost equal to $V^\pi(x) + w_t^\pi(x)$, where $w_t^\pi(x)$ is a zero-mean noise. We then update $V_t^\pi, \pi \in \Delta$, (synchronously at each state) by

$$V_{t+1}^\pi(x) = (1-\alpha_t) V_t^\pi(x) + \alpha_t \left(V^\pi(x) + w_t^\pi(x) \right), x \in X, \quad (4)$$

where α_t is a deterministic scalar stepsize parameter. Note that $E[w_t^\pi(x) | \mathcal{F}_t] = 0$ for any $x \in X$, where $\mathcal{F}_t$ denotes the history of the algorithm up to and including the point where $V_t^\pi, \pi \in \Delta$, has become available, but before simulating the sample paths that will determine the next update. $w_t^\pi(x)$ is a function of the random variables contained in $\mathcal{F}_{t+1}$. The variance of $w_t^\pi(x)$ conditioned on $\mathcal{F}_t$ is bounded by some constant because there are finitely many policies and states. Furthermore, for each $\pi \in \Delta$, $w_t^\pi(x)$ over t are independent identically distributed random variables.

The function V_0^π can be set to be an arbitrary function in $B(X)$. Because the cumulative discounted cost for any sample path is bounded such that $\max_{x \in X} |V^\pi(x) + w_t^\pi(x)| \leq \max_{x \in X, a \in A} |C(x,a)|/(1-\gamma)$, we better choose V_0^π such that $|V_0^\pi(x)| \leq \max_{x \in X, a \in A} |C(x,a)|/(1-\gamma)$. In the special case where $V_0^\pi(x) = 0, x \in X$ for all $\pi \in \Delta$ and $\alpha_t = 1/(t+1)$, each $V_t^\pi(x), \pi \in \Delta$, is equal to the average of the observed cumulative costs of t independently generated sample paths that start at x, and converges to $V^\pi(x)$ for all $x \in X$ as $t \to \infty$ and if we apply the multi-policy improvement (1) *after* the convergence, we would have the original parallel rollout method.

This stochastic iterative algorithm is "conceptual" because we cannot simulate over infinite horizon in practice. As in the remark given in Section 6 in [6], we can simulate each policy over a finite but "long" horizon to obtain the infinite horizon trajectory cost. As the horizon size increases, due to the effect of the discount factor, the finite horizon approximation becomes a very close approximation to the infinite horizon trajectory cost. Indeed, the effectiveness of the parallel rollout policy implemented with the average of the observed cumulative costs of a fixed number of independently generated sample paths over a finite horizon has been shown in the context of on-line control for MDPs in [2].

Alternatively, we can consider only problems with a zero-cost absorbing state or letting the simulation process terminate with probability γ at each stage, and to accumulate undiscounted costs [6].

Theorem 1. Assume that

$$\sum_{t=0}^{\infty} \alpha_t = \infty, \qquad \sum_{t=0}^{\infty} \alpha_t^2 < \infty.$$

Then, for any $x \in X$,

$$\lim_{t \to \infty} V^{\mu_t^{\text{pr}}}(x) \leq \min_{\pi \in \Delta} V^\pi(x).$$

We will use the following lemma at several places for the proofs throughout the present paper:

Lemma 2.1. Given $\pi \in \Pi$ and $\tau \in \mathbb{R}$, suppose that there exists $\Phi \in B(X)$ for which

$$T_\pi(\Phi)(x) \leq \Phi(x) + \tau, x \in X. \quad (5)$$

Then, $V^\pi(x) \leq \Phi(x) + \frac{\tau}{1-\gamma}$ for all $x \in X$.

Proof: By successive applications of the T_π-operator to both sides of (5) and the monotonicity property of the operator, we have that for all $x \in X$,

$$\lim_{n \to \infty} T_\pi^n(\Phi)(x) \leq \Phi(x) + \lim_{n \to \infty} \tau(1 + \gamma + \gamma^2 + \cdots + \gamma^{n-1}).$$

It is well-known that T_π is a contraction mapping in $B(X)$ and that iterative application of T_π on any initial value function converges monotonically to the fixed point V^π. Therefore, $\lim_{n \to \infty} T_\pi^n(\Phi)(x) = V^\pi(x), x \in X$, which proves the lemma. ∎

Proof of Theorem 1: The parts of the proof follows the proof idea of Proposition 1 in [6].

For any $x \in X$ and any $\pi \in \Delta$,

$$\begin{aligned}
&T_\pi(V_{t+1}^\pi)(x) \\
&= C(x, \pi(x)) + \gamma \sum_{y \in X} P_{xy}^{\pi(x)} [(1-\alpha_t) V_t^\pi(y) \\
&\quad + \alpha_t (V^\pi(y) + w_t^\pi(y))] \\
&= C(x, \pi(x)) + (1-\alpha_t) \gamma \sum_{y \in X} P_{xy}^{\pi(x)} V_t^\pi(y) \\
&\quad + \alpha_t \gamma \sum_{y \in X} P_{xy}^{\pi(x)} \left(V^\pi(y) + w_t^\pi(y) \right) \\
&= T_\pi(V_t^\pi)(x) - \alpha_t \gamma \sum_{y \in X} P_{xy}^{\pi(x)} V_t^\pi(y) - \alpha_t C(x, \pi(x)) \\
&\quad + \alpha_t C(x, \pi(x)) + \alpha_t \gamma \sum_{y \in X} P_{xy}^{\pi(x)} \left(V^\pi(y) + w_t^\pi(y) \right) \\
&= T_\pi(V_t^\pi)(x) - \alpha_t T_\pi(V_t^\pi) + \alpha_t T_\pi(V^\pi)(x)
\end{aligned}$$

$$
\begin{aligned}
&+\alpha_t\gamma\sum_{y\in X}P_{xy}^{\pi(x)}w_t^\pi(y)\\
&=(1-\alpha_t)T_\pi(V_t^\pi)(x)+\alpha_tV^\pi(x)+\alpha_t\gamma\sum_{y\in X}P_{xy}^{\pi(x)}w_t^\pi(y)\\
&=(1-\alpha_t)\left(T_\pi(V_t^\pi)(x)-V_t^\pi(x)\right)+(1-\alpha_t)V_t^\pi(x)\\
&\quad+\alpha_t(V^\pi(x)+w_t^\pi(x))+\alpha_t\gamma\sum_{y\in X}P_{xy}^{\pi(x)}w_t^\pi(y)-\alpha_tw_t^\pi(x)\\
&=V_{t+1}^\pi(x)+(1-\alpha_t)\left(T_\pi(V_t^\pi)(x)-V_t^\pi(x)\right)\\
&\quad+\alpha_t\left(\gamma\sum_{y\in X}P_{xy}^{\pi(x)}w_t^\pi(y)-w_t^\pi(x)\right).
\end{aligned}
$$

We have established that for any $x\in X$ and any $\pi\in\Delta$, $T_\pi(V_{t+1}^\pi)(x)-V_{t+1}^\pi(x)=(1-\alpha_t)(T_\pi(V_t^\pi)(x)-V_t^\pi(x))+\alpha_t\eta_t^\pi(x)$, where $\eta_t^\pi(x)=\gamma\sum_{y\in X}P_{xy}^{\pi(x)}w_t^\pi(y)-w_t^\pi(x)$. Let $X_t^\pi(x)=T_\pi(V_t^\pi)(x)-V_t^\pi(x), x\in X$. Then from the standard results on convergence of stochastic approximations [1, Chapter 4], $X_t^\pi(x)$ updated with

$$X_{t+1}^\pi(x)=(1-\alpha_t)X_t^\pi(x)+\alpha_t\eta_t^\pi(x)$$

converges to zero with probability one. Therefore, we have that w.p. 1 (with probability 1) for any $x\in X$ and any $\pi\in\Delta$,

$$\lim_{t\to\infty}T_\pi(V_t^\pi)(x)-V_t^\pi(x)=0,$$

which implies that for every $\epsilon>0$, there exists a time $t(\epsilon)$ such that for all $x\in X$ and $\pi\in\Delta$,

$$|T_\pi(V_t^\pi)(x)-V_t^\pi(x)|\le\epsilon,\forall\ t\ge t(\epsilon). \tag{6}$$

Because of the monotonicity of T_π-operator and $J_t(x)=\min_{\pi\in\Delta}V_t^\pi(x), x\in X$,

$$T_\pi(J_t)(x)\le T_\pi(V_t^\pi)(x), x\in X,\pi\in\Delta$$

and

$$T(J_t)(x)\le T_\pi(J_t)(x), x\in X,\pi\in\Delta.$$

Therefore, for any $x\in X$,

$$T(J_t)(x)-\min_{\pi\in\Delta}V_t^\pi(x)\le\epsilon,\forall\ t\ge t(\epsilon).$$

We have that for every $\epsilon>0$, there exists a time $t(\epsilon)$ such that

$$\max_{x\in X}\left(T(J_t)(x)-J_t(x)\right)\le\epsilon,\forall\ t\ge t(\epsilon). \tag{7}$$

From the definition of $\mu_t^{\rm pr}$, $T_{\mu_t^{\rm pr}}(J_t)(x)=T(J_t)(x), x\in X$, we have that

$$T_{\mu_t^{\rm pr}}(J_t)(x)\le J_t(x)+\max_{x\in X}\left(T(J_t)(x)-J_t(x)\right), x\in X,$$

which implies that by Lemma 2.1 and (7), for any $x\in X$,

$$
\begin{aligned}
V^{\mu_t^{\rm pr}}(x)&\le J_t(x)+\frac{\max_{x\in X}\left(T(J_t)(x)-J_t(x)\right)}{1-\gamma}\\
&\le J_t(x)+\frac{\epsilon}{1-\gamma},\forall\ t\ge t(\epsilon).
\end{aligned} \tag{8}
$$

From (6), for any $\pi\in\Delta$, $V_t^\pi(x)-T_\pi(V_t^\pi)(x)\le\epsilon$ for all $t\ge t(\epsilon)$ so that by Lemma 2.1

$$V_t^\pi(x)\le V^\pi(x)+\frac{\epsilon}{1-\gamma}, x\in X,\forall\ t\ge t(\epsilon).$$

Therefore,

$$\min_{\pi\in\Delta}V_t^\pi(x)\le\min_{\pi\in\Delta}V^\pi(x)+\frac{\epsilon}{1-\gamma}, x\in X,\forall\ t\ge t(\epsilon). \tag{9}$$

Combining (8) and (9),

$$V^{\mu_t^{\rm pr}}(x)\le\min_{\pi\in\Delta}V^\pi(x)+\frac{2\epsilon}{1-\gamma}, x\in X,\forall\ t\ge t(\epsilon). \tag{10}$$

Because we can choose $t(\epsilon)$ for any $\epsilon>0$ arbitrarily close to zero, we have the desired convergence. □

The algorithm we studied generates a sample path from every initial state at each iteration. We can choose a single state x, randomly, uniformly, and independently from everything else, and generate a single path starting from x. We then update $V_t^\pi(x),\pi\in\Delta$, only and make no change on $V_t^\pi(x'), x'\ne x$. This variant of the previously presented algorithm also converges uniformly over X in the sense that $\lim_{t\to\infty}V^{\mu_t^{\rm pr}}(x)\le\min_{\pi\in\Delta}V^\pi(x), x\in X$. See the related discussion in Section 3 in [6]. We also remark that the iterative computation given as

$$\tilde V_{t+1}(x):=(1-\alpha_t)\tilde V_t(x)+\alpha_t(V^{\mu_t^{\rm pr}}(x)+w^{\mu_t^{\rm pr}}(x)), x\in X,$$

converges such that

$$\limsup_{t\to\infty}\tilde V_t(x)\le\min_{\pi\in\Delta}V^\pi(x), x\in X.$$

We skip the details as the arguments are straightforward applications of the parts in the proof of Theorem 1. The following corollary on the finite-time bound can also be stated with the slight change of the proof of Theorem 1:

Corollary 2.1. For any $t\ge0$ and $x\in X$, $V^{\mu_t^{\rm pr}}(x)\le\min_{\pi\in\Delta}V^\pi(x)+2\max_{\pi\in\Delta,x\in X}|T_\pi(V_t^\pi)(x)-V_t^\pi(x)|/(1-\gamma)$.

TD(λ)-based policy evaluation The previous section discussed a stochastic iterative parallel-rollout computation method where the cumulative cost of a sample path is obtained by Monte Carlo simulation. We extend the result of the previous section into the case where the well-known TD(λ) algorithm [5] is used for policy evaluation, instead of Monte Carlo simulation as Tsitsiklis has done for optimistic PI [6].

The stochastic iterative computation of parallel rollout with TD(λ) is the same to the previous description except that (4) is replaced by the following: for $\pi\in\Delta$,

$$
\begin{aligned}
&V_{t+1}^\pi(x)=\\
&\quad(1-\alpha_t)V_t^\pi(x)+\alpha_t(1-\lambda)\Theta(x)+\alpha_tw_t^\pi(x), x\in X
\end{aligned} \tag{11}
$$

where the value function Θ defined over X is obtained by

$$\Theta(x)=\left(\sum_{k=0}^\infty\lambda^kT_\pi^{k+1}(V_t^\pi)\right)(x), x\in X$$

and $\lambda\in[0,1)$ (the $\lambda=1$ case corresponds to the update rule of the previous section). Note that we are abusing the notation for the noise; here $w_t^\pi(x)$ is a zero-mean noise that reflects the difference between the observed "temporal differences" and their expected values.

Theorem 2. Assume that

$$\sum_{t=0}^\infty\alpha_t=\infty,\qquad\sum_{t=0}^\infty\alpha_t^2<\infty.$$

Then, for any $x \in X$,

$$\lim_{t\to\infty} V^{\mu_t^{\rm pr}}(x) \le \min_{\pi\in\Delta} V^{\pi}(x).$$

Proof: Let $\beta_t^\pi = \max_{x\in X} |T_\pi(V_t^\pi)(x) - V_t^\pi(x)|, \pi \in \Delta$, and $e \in \mathbb{R}^{|X|}$ be the vector with all components equal to 1. For any $x \in X$ and any $\pi \in \Delta$,

$$\begin{aligned}
&T_\pi(V_{t+1}^\pi)(x)\\
&= (1-\alpha_t)T_\pi(V_t^\pi)(x) + \alpha_t(1-\lambda)T_\pi(\sum_{k=0}^{\infty}\lambda^k T_\pi^{k+1}(V_t^\pi))(x)\\
&\quad +\alpha_t\gamma\sum_{y\in X} P_{xy}^{\pi(x)} w_t^\pi(y)\\
&= (1-\alpha_t)V_t^\pi(x) + (1-\alpha_t)(T_\pi(V_t^\pi)(x) - V_t^\pi(x))\\
&\quad +\alpha_t(1-\lambda)\left(\sum_{k=0}^{\infty}\lambda^k T_\pi^{k+1}(T_\pi(V_t^\pi))\right)(x)\\
&\quad +\alpha_t\gamma\sum_{y\in X} P_{xy}^{\pi(x)} w_t^\pi(y)\\
&\le (1-\alpha_t)V_t^\pi(x) + (1-\alpha_t)(T_\pi(V_t^\pi)(x) - V_t^\pi(x))\\
&\quad +\alpha_t(1-\lambda)\left(\sum_{k=0}^{\infty}\lambda^k T_\pi^{k+1}(V_t^\pi + \beta_t^\pi e)\right)(x)\\
&\quad +\alpha_t\gamma\sum_{y\in X} P_{xy}^{\pi(x)} w_t^\pi(y)\\
&= V_{t+1}^\pi(x) + (1-\alpha_t)(T_\pi(V_t^\pi)(x) - V_t^\pi(x))\\
&\quad +\alpha_t(1-\lambda)\sum_{k=0}^{\infty}\lambda^k\gamma^{k+1}\beta_t^\pi + \alpha_t\eta_t^\pi(x)\\
&\le V_{t+1}^\pi(x) + (1-\alpha_t)(T_\pi(V_t^\pi)(x) - V_t^\pi(x))\\
&\quad +\alpha_t\gamma\beta_t^\pi + \alpha_t\eta_t^\pi(x), \qquad (12)
\end{aligned}$$

where $\eta_t^\pi(x) = \gamma\sum_{y\in X} P_{xy}^{\pi(x)} w_t^\pi(y) - w_t^\pi(x)$. Let $X_t^\pi(x) = T_\pi(V_t^\pi)(x) - V_t^\pi(x), x \in X$. Then we have that

$$X_{t+1}^\pi(x) \le (1-\alpha_t)X_t^\pi(x) + \alpha_t\gamma\max_{x\in X}|X_t^\pi(x)| + \alpha_t\eta_t^\pi(x).$$

Then comparing $X_t^\pi(x)$ with the sequence $Y_t(x)$ defined by $Y_0(x) = X_0^\pi(x)$ and

$$Y_{t+1}(x) = (1-\alpha_t)Y_t(x) + \alpha_t\gamma\max_{x\in X}|Y_t(x)| + \alpha_t\eta_t^\pi(x),$$

and noting that $X_t^\pi(x) \le Y_t(x)$ for all t and the mapping $V \to \gamma e\max_{x\in X}|V(x)|$ for $V \in B(X)$ is a maximum norm contraction (see, [6, p.70]), w.p. 1

$$\limsup_{t\to\infty} X_t^\pi(x) \le 0, x \in X.$$

Therefore, for every $\epsilon > 0$, there exists a time $t(\epsilon)$ such that for all $x \in X$ and $\pi \in \Delta$,

$$T_\pi(V_t^\pi)(x) - V_t^\pi(x) \le \epsilon, \forall\ t \ge t(\epsilon). \qquad (13)$$

With the same reasoning in the proof of Theorem 1, from (13) we have that for every $\epsilon > 0$, there exists a time $t(\epsilon)$ such that

$$\max_{x\in X}(T(J_t)(x) - J_t(x)) \le \epsilon, \forall\ t \ge t(\epsilon). \qquad (14)$$

Therefore,

$$V^{\mu_t^{\rm pr}}(x) \le J_t(x) + \frac{\epsilon}{1-\gamma}, x \in X, \forall\ t \ge t(\epsilon). \qquad (15)$$

Using the relationship of $T_\pi(V_t^\pi)(x) - V_t^\pi(x) \ge -\beta_t^\pi, x \in X$, we can establish $T_\pi(V_{t+1}^\pi)(x) \ge V_{t+1}^\pi(x) + (1-\alpha_t)(T_\pi(V_t^\pi)(x) - V_t^\pi(x)) - \alpha_t\gamma\beta_t^\pi + \alpha_t\eta_t^\pi(x)$, making w.p. 1,

$$\liminf_{t\to\infty} X_t^\pi(x) \ge 0, x \in X.$$

Therefore, for every $\delta > 0$, there exists a time $t(\delta)$ such that $V_t^\pi(x) - T_\pi(V_t^\pi)(x) \le \delta$ for all $t \ge t(\delta)$. We have that for any $\pi \in \Delta$,

$$V_t^\pi(x) - V^\pi(x) \le \frac{\delta}{1-\gamma}, x \in X, \forall\ t \ge t(\delta).$$

This together with (15) implies that $\forall\ t \ge \max\{t(\epsilon), t(\delta)\}$,

$$V^{\mu_t^{\rm pr}}(x) \le \min_{\pi\in\Delta} V^\pi(x) + \frac{2\max\{\epsilon,\delta\}}{1-\gamma}, x \in X. \qquad (16)$$

Because we can choose $\max\{t(\epsilon), t(\delta)\}$ for any $\epsilon, \delta > 0$ arbitrarily close to zero, we have the desired convergence. ∎

2.2 Policy Switching

The stochastic iterative computation for the parallel rollout policy requires knowing the model of the given MDP M. To compute $\mu_t^{\rm pr}$ with respect to J_t in (3), we need to know $C(x,a), x \in X, a \in A$ and $P_{xy}^a, x, y \in X, a \in A$. On the other hand, the computation of the policy switching policy below can be done in *model-free* environment as long as Monte Carlo policy evaluation is done in a model-free manner. We only discuss the case of Monte Carlo policy evaluation here as the case of TD(λ)-based policy iteration can be studied by a similar reasoning given below with the proof idea of Theorem 2.

We use the same notations used in the parallel rollout case. At each iteration $t \ge 0$, we have available value functions $V_t^\pi, \pi \in \Delta$, defined over X and let $\mu_t^{\rm ps} \in \Pi$ be a policy such that

$$\mu_t^{\rm ps}(x) \in \arg\min_{\pi\in\Delta}(V_t^\pi(x))(x), x \in X. \qquad (17)$$

Computation of V_t^π is the same as before: for each policy $\pi \in \Delta$ starting with each state $x \in X$, we generate a corresponding sample path over infinite horizon starting with state x then update $V_t^\pi, \pi \in \Delta$, (synchronously at each state) by

$$V_{t+1}^\pi(x) = (1-\alpha_t)V_t^\pi(x) + \alpha_t\left(V^\pi(x) + w_t^\pi(x)\right), \qquad (18)$$

$x \in X$, where α_t is again a deterministic scalar stepsize parameter.

For the convergence analysis, we define a value function J_t as $J_t(x) = \min_{\pi\in\Delta} V_t^\pi(x), x \in X$.

Theorem 3. Assume that

$$\sum_{t=0}^{\infty}\alpha_t = \infty, \qquad \sum_{t=0}^{\infty}\alpha_t^2 < \infty.$$

Then, for any $x \in X$,

$$\lim_{t\to\infty} V^{\mu_t^{\rm ps}}(x) \le \min_{\pi\in\Delta} V^\pi(x).$$

Proof: From the proof of Theorem 1, for every $\epsilon > 0$, there exists a time $t(\epsilon)$ such that for all $x \in X$ and $\pi \in \Delta$,

$$|T_\pi(V_t^\pi)(x) - V_t^\pi(x)| \le \epsilon, x \in X, \forall\ t \ge t(\epsilon). \qquad (19)$$

Therefore, by Lemma 2.1, for any $\pi \in \Delta$,

$$|V^{\pi}(x) - V_t^{\pi}(x)| \leq \frac{\epsilon}{1-\gamma}, x \in X, \forall\ t \geq t(\epsilon). \qquad (20)$$

From the definition of $\mu_t^{\rm ps}$, for a given $x \in X$, $\mu_t^{\rm ps}(x) = \pi'(x)$ for some $\pi' \in \Delta$ such that $V_t^{\pi'}(x) \leq V_t^{\pi}(x)$. Now,

$$\begin{aligned}
&T_{\mu_t^{\rm ps}}(J_t)(x) \\
&= C(x, \mu_t^{\rm ps}(x)) + \gamma \sum_{y \in X} P_{xy}^{\mu_t^{\rm ps}(x)} J_t(y) \\
&= C(x, \pi'(x)) + \gamma \sum_{y \in X} P_{xy}^{\pi'(x)} J_t(y) \\
&\leq C(x, \pi'(x)) + \gamma \sum_{y \in X} P_{xy}^{\pi'(x)} V_t^{\pi'}(y) \\
&= C(x, \pi'(x)) + \gamma \sum_{y \in X} P_{xy}^{\pi'(x)} \left(V^{\pi'}(y) + V_t^{\pi'}(y) - V^{\pi'}(y) \right) \\
&\leq T_{\pi'}(V^{\pi'})(x) + \gamma \max_{x \in X} |V_t^{\pi'}(x) - V^{\pi'}(x)| \\
&\leq V^{\pi'}(x) + \frac{\gamma\epsilon}{1-\gamma}, \forall\ t \geq t(\epsilon) \\
&\leq V^{\pi'}(x) - V_t^{\pi'}(x) + V_t^{\pi'}(x) + \frac{\gamma\epsilon}{1-\gamma}, \forall\ t \geq t(\epsilon) \\
&\leq \frac{\epsilon}{1-\gamma} + J_t(x) + \frac{\gamma\epsilon}{1-\gamma}, \forall\ t \geq t(\epsilon),
\end{aligned}$$

where the first term is from (20) and the second term is from the definition of J_t, i.e., $V_t^{\pi'}(x) = J_t(x), x \in X$. Therefore,

$$T_{\mu_t^{\rm ps}}(J_t)(x) \leq J_t(x) + \frac{\epsilon(1+\gamma)}{1-\gamma}, x \in X, \forall\ t \geq t(\epsilon).$$

Applying Lemma 2.1,

$$V^{\mu_t^{\rm ps}}(x) \leq \min_{\pi \in \Delta} V^{\pi}(x) + \frac{\epsilon(1+\gamma)}{(1-\gamma)^2}, x \in X, \forall\ t \geq t(\epsilon).$$

Because we can choose $t(\epsilon)$ for any $\epsilon > 0$ arbitrarily close to zero, we have the desired convergence. ∎

Corollary 2.2. For any $t \geq 0$ and $x \in X$,

$$\begin{aligned}
V^{\mu_t^{\rm ps}}(x) &\leq \min_{\pi \in \Delta} V^{\pi}(x) \\
&+ \frac{(1+\gamma)}{(1-\gamma)^2} \cdot \max_{\pi \in \Delta, x \in X} |T_{\pi}(V_t^{\pi})(x) - V_t^{\pi}(x)|. \qquad (21)
\end{aligned}$$

3. CONCLUDING REMARKS

The algorithms presented in this paper is in the context of "off-line" computation. In practice, we would apply the stochastic iterative computation methods for a finite number of iterations with a finite-horizon sample path simulation. However, suppose that we want to apply the methods in on-line manner. This can be done as follows: at the current state $x_t \in X$ at time t, we asynchronously update only the x_t-component of the V_t^{π}-function for each $\pi \in \Delta$ in (4) and (18). We then compute the parallel rollout policy $\mu_t^{\rm pr}$ (or the policy switching policy $\mu_t^{\rm ps}$) and apply the action prescribed by the computed policy to the system for on-line control. In particular, policy switching is very attractive because it escapes from the requirement of knowing the model and also from the curse of the dimensionality problem in solving MDPs.

Under the assumption that the given MDP M is communicating, i.e., any state can be reached from any other state for any Markov chain induced from fixing any policy in Π, each state $x \in X$ is visited infinitely often with probability 1. If the assumption holds, even though the V_t^{π}-function for each $\pi \in \Delta$ is updated asynchronously only at the currently visited state, after a "long" time, $V_t^{\pi}(x)$ closely estimates $V^{\pi}(x)$ for all $x \in X$, i.e., there exists a time $t(\epsilon)$ such that $\max_{x \in X}(T(J_t)(x) - J_t(x)) \leq \epsilon$. Therefore, $\mu_t^{\rm pr}$ (similarly, $\mu_t^{\rm ps}$) eventually converges uniformly over X in the sense that $\lim_{t \to \infty} V^{\mu_t^{\rm pr}}(x) \leq \min_{\pi \in \Delta} V^{\pi}(x), x \in X$.

REFERENCES

[1] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*. Athena Scientific, Belmont, Massachusetts, 1996.

[2] H. S. Chang, R. Givan, and E. K. P. Chong, "Parallel rollout for on-line solution of partially observable Markov decision processes," *Discrete Event Dynamic Systems: Theory and Application*, vol. 14, no. 3, 2004, pp. 309–341.

[3] H. S. Chang and S. I. Marcus, "Approximate receding Horizon approach for Markov decision processes: average reward case," *J. of Mathematical Analysis and Applications*, vol. 286, 2003, pp. 636–651.

[4] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, New York, 1994.

[5] R. Sutton and A. Barto, *Reinforcement Learning*. MIT Press, 2000.

[6] J. N. Tsitsiklis, "On the convergence of optimistic policy iteration," *J. of Machine Learning Research*, vol. 3, 2002, pp. 59–72.