Proceedings of the
44th IEEE Conference on Decision and Control, and
the European Control Conference 2005
Seville, Spain, December 12-15, 2005

On Solving Controlled Markov Set-Chains via Multi-Policy Improvement

Hyeong Soo Chang*
Department of Computer Science and Engineering
Sogang University
Seoul, Korea
hschang@sogang.ac.kr

Edwin K. P. Chong**
Department of Electrical and Computer Engineering
Colorado State University
Fort Collins, CO 80523-1373, USA
echong@engr.colostate.edu

Abstract— **We present formal methods of improving multiple policies for solving controlled Markov set-chains with infinite-horizon discounted reward criteria. The multi-policy improvement methods follow the spirit of parallel rollout for solving Markov decision processes (MDPs). In particular, these methods are useful for on-line control of Markov set-chains and for approximately solving MDPs via state aggregation. We further discuss issues on designing a policy-iteration type algorithm based on our policy improvement methods.**

I. Introduction

When dealing with a Markov Decision Process (MDP) [6], the state-transition probability matrix typically is assumed to be uniquely given at each decision time. The probabilistic state transition at each time is uniquely determined from a given transition probability distribution associated with the current state and the current action, whether the distribution is stationary or non-stationary or whether it is known to a decision maker or not.

Kurano *et al.* [5] extend the usual MDP model to the case where the transition probability varies in some given domain at each decision time, and its variation is unobservable or unknown (see, e.g., [5] for example problems). In doing so, they develop a novel model called a "controlled Markov set-chain," based on Markov set-chains [3], and study an optimal control problem with a total expected discounted reward criterion under some partial order. In their generalization of the MDP model, each state and action pair is associated with a range of probability distributions and, at each decision time, a probability distribution is arbitrarily selected by some unknown adversary of the system. In this view, the controlled Markov set-chain model is suitable for designing a robust controller. Furthermore, the model can be used for sensitivity analysis of an MDP where the transition probability parameters are perturbed. The model also can be used as an approximate-solution approach for solving large MDPs via aggregation of the states. By grouping the states in a given MDP, we can induce a controlled Markov set-chain model with a much smaller state space (see [2] for a related discussion).

Based on appropriately defined contraction operators, Kurano *et al.* [5] establish an optimality equation satisfied by an optimal policy (optimal in a certain partial-order sense), and some results that induce a value-iteration [6] type algorithm for solving problems modeled by controlled Markov set-chains. A condition for policy improvement [6] for a *single* policy is provided, but no policy-iteration (PI) type algorithm based on the condition is discussed explicitly in their paper.

In this paper, we present formal methods for improving *multiple* policies in a partial-order sense and discuss some issues on designing a PI-type algorithm for controlled Markov set-chains. Our multi-policy improvement methods follow the spirit of parallel rollout [1] for solving MDPs. Along the same line, they are useful for on-line control of Markov set-chains and for approximately solving MDPs via state aggregation. We further discuss issues on designing a policy-iteration type algorithm based on our policy improvement methods.

Some related models based on MDPs have been studied in the operations-research literature by White and Eldeib [8], and Satia and Lave [7], under the rubric of MDPs with "imprecisely known transition probabilities." Related work has also been reported in the artificial-intelligence literature by Givan *et al.* [2], who discuss "bounded parameter Markov Decision Processes (BMDPs)." The controlled Markov set-chain model is very similar to BMDPs. See Section 8 in [2] for a discussion on the relationship of BMDPs with other models. Kalyanasundaram *et al.* [4] study continuous-time MDPs with unknown transition rates and average reward criteria, and develop a PI-type algorithm based on single-policy improvement, for obtaining robust ("max-min") policies.

This paper is organized as follows. In Section II, we formally describe the controlled Markov set-chain model and some preliminaries. In Section III, we then discuss two versions of the multi-policy improvement method, and the main theoretical results. We also discuss the design of a PI-type algorithm for solving controlled Markov set-chains. In Section IV, we briefly describe the use of our multi-policy improvement results. We conclude the present paper in Section V.

*H. S. Chang is also affiliated with Program of Integrated Biotechnology at Sogang University. This work was supported by the Ministry of Science and Technology 21st Century Frontier Program: Intelligent Robot Project in 2005.

**This work was supported in part by NSF through grant ANI-0207892.

0-7803-9568-9/05/$20.00 ©2005 IEEE

II. Controlled Markov Set-Chains

In this section we provide a formal description of controlled Markov set-chains, following the notation of [5] (see [5] for more detailed discussion). A controlled Markov set-chain model is a four-tuple $M = (X, A, R, P = \langle \underline{p}, \overline{p} \rangle)$, where X is a finite set of states, A is a finite set of actions, $R : X \times A \rightarrow \mathbb{R}^+$ represents a bounded nonnegative reward function, and $P = \langle \underline{p}, \overline{p} \rangle$ is an "interval transition function." To elaborate, let $\mathbb{R}_+^{1\times|X|}$ denote the set of vectors in $\mathbb{R}^{1\times|X|}$ with entrywise nonnegative elements. Then, for each $(x, a) \in X \times A$, $\underline{p} = \underline{p}(\cdot|x, a) \in \mathbb{R}_+^{1\times|X|}$ and $\overline{p} = \overline{p}(\cdot|x, a) \in \mathbb{R}_+^{1\times|X|}$ with $\underline{p} \leq \overline{p}$ (the relations $\leq$, $<$, and $=$ used in a vector or a matrix context are defined componentwise throughout the present paper). We assume that $\langle \underline{p}, \overline{p} \rangle := \{p | p \text{ is a probability distribution over } X \text{ with } \underline{p} \leq p \leq \overline{p}\}$ is nonempty.

If the system is in state $x \in X$ and an action $a \in A$ is taken at x, then the system makes a transition from state x to a random next-state $y \in X$ according the probability $p(y|x, a)$, and a reward of $R(x, a)$ is obtained. This process is repeated at the state y. To decide what action a to take at each state x, the decision maker only knows that $p(\cdot|x, a)$ is arbitrarily selected from $\langle \underline{p}(\cdot|x, a), \overline{p}(\cdot|x, a) \rangle$. The decision maker wishes to maximize the total expected discounted reward over an infinite horizon. Note that if $\underline{p} = \overline{p}$, then M reduces to a standard MDP [6].

We define a stationary policy π as a mapping from X to A, and let Π be the set of all possible policies. We associate with each $\pi \in \Pi$ the $|X|$-dimensional column vector $R(\pi) \in \mathbb{R}_+^{|X|}$, where the entry of $R(\pi)$ associated with state x is $R(x, \pi(x))$. Consider the set of stochastic matrices $\mathcal{P}(\pi) := \langle \underline{P}, \overline{P} \rangle$, where

$$\langle \underline{P}, \overline{P} \rangle = \{P | P \text{ is a stochastic matrix with } \underline{P} \leq P \leq \overline{P}\},$$

and the rows associated with state x for $\underline{P}$ and $\overline{P}$ are $\underline{p}(\cdot|x, \pi(x))$ and $\overline{p}(\cdot|x, \pi(x))$, respectively.

We will need some more notation involving sets of stochastic matrices. Let

$$\mathcal{M}_{m\times n} = \{\mathcal{A} = \langle \underline{A}, \overline{A} \rangle | \underline{A} \leq \overline{A}, \underline{A}, \overline{A} \in \mathbb{R}_+^{m\times n}\}.$$

Identifying any stochastic matrix P with the set $\langle P, P \rangle = \{P\}$, any state transition matrix is associated with some element of $\mathcal{M}_{|X|\times|X|}$. The product of $\mathcal{A}$ and $\mathcal{B}$ in $\mathcal{M}_{|X|\times|X|}$ is defined by

$$\mathcal{A}\mathcal{B} = \{AB | A \in \mathcal{A}, B \in \mathcal{B}\}$$

(the product of the elements in $\mathcal{M}_{|X|\times 1}$ and $\mathcal{M}_{|X|\times|X|}$ is also similarly defined). For any sequence $\{\mathcal{A}_i\}_{i=1}^{\infty}$ with $\mathcal{A}_i \in \mathcal{M}_{|X|\times|X|}$, $i \geq 1$, the *multiproduct* is defined inductively by

$$\mathcal{A}_1\mathcal{A}_2 \cdots \mathcal{A}_k := (\mathcal{A}_1 \cdots \mathcal{A}_{k-1})\mathcal{A}_k, k \geq 2.$$

Similarly, for any vector $v \in \mathbb{R}^{|X|}$ and $\mathcal{A} \in \mathcal{M}_{|X|\times|X|}$, $v\mathcal{A} = \{vA | A \in \mathcal{A}\}$.

For a given policy $\pi \in \Pi$, we define *the value of following* π with an initial state $x \in X$ as

$$V^{\pi}(x) = e_x \left\{ \sum_{t=0}^{\infty} \gamma^t \left(\prod_{i=0}^{t} P_i \right) R(\pi) \middle| P_i \in \mathcal{P}(\pi), \right. \\ \left. i \geq 1, P_0 = I \right\},$$

where $\gamma \in (0, 1)$ is a discount factor, $e_x \in \mathbb{R}_+^{|X|}$ is the unit vector whose entry associated with x is 1 and all of the other entries are zero, and I denotes the identity matrix. (See also Lemma 2.1 in [5] for more on this definition.)

Denote the set of all bounded and closed intervals in $\mathbb{R}^+$ by $C(\mathbb{R}^+)$. Next, denote the set of all $|X|$-dimensional column vectors whose elements are in $C(\mathbb{R}^+)$ by $C(\mathbb{R}_+^{|X|})$. Note that given $\pi \in \Pi$, we have $V^{\pi}(x) \in C(\mathbb{R}^+)$, $x \in X$, and the value function V^{π} is in $C(\mathbb{R}_+^{|X|})$. Hence, we can write $V^{\pi} = [\underline{V}^{\pi}, \overline{V}^{\pi}]$ with $\underline{V}^{\pi}, \overline{V}^{\pi} \in \mathbb{R}_+^{|X|}$ and $\underline{V}^{\pi} \leq \overline{V}^{\pi}$.

We now define a *partial order* $(\geq, >)$ on $C(\mathbb{R}^+)$: for $[c_1, c_2]$ and $[d_1, d_2]$ in $C(\mathbb{R}^+)$, we write $[c_1, c_2] \geq [d_1, d_2]$ if $c_1 \geq d_1$ and $c_2 \geq d_2$, and $[c_1, c_2] > [d_1, d_2]$ if $[c_1, c_2] \geq [d_1, d_2]$ and $[c_1, c_2] \neq [d_1, d_2]$.

Using the above partial order, we then say that a *policy* π^* *is optimal in* Π if there does not exist $\pi \in \Pi$ such that $V^{\pi^*} < V^{\pi}$. Our goal for a given M is to find an optimal policy $\pi^* \in \Pi$. Note that the optimal policy may not be unique, and, for two optimal policies π_1^* and π_2^*, the value of following each policy may be different, i.e., $V^{\pi_1^*}$ and $V^{\pi_2^*}$ are not necessarily equal. This is in contrast to the uniqueness of the optimal value function in standard MDPs.

Kurano *et al.* [5] prove the existence of an optimal stationary policy π^* and establish an optimality equation uniquely satisfied by the policy's value function V^{π^*}. They also provide some results that induce a value-iteration type algorithm [6] to compute V^{π^*} by defining relevant contraction operators (thereby obtaining π^*). They further provide a sufficient condition for single-policy improvement (see Corollary 4.1 and 4.2 in [5] and Lemma 1 below), but they do not explicitly discuss any policy improvement method; specifically, they do not provide any type of policy-iteration (PI) algorithm in their paper.

III. Multi-Policy Improvement

The policy improvement step of any PI-type algorithm for solving MDPs is based on improving a single policy. Recently, Chang *et al.* [1] extend this single-policy improvement method to a multi-policy improvement method using an approach called "parallel rollout." Applying a similar policy improvement step to the case of controlled Markov set-chains is nontrivial, because in the latter case we have to deal with simultaneously improving the *intervals* associated with each state (where "improvement" is in the sense of the partial order). Our main goal here is to provide some useful results regarding multi-policy improvement for controlled Markov set-chains.

For each $(x, a) \in X \times A$, let

$$p(x, a) := \langle \underline{p}(\cdot|x, a), \overline{p}(\cdot|x, a) \rangle.$$

For each policy $\pi \in \Pi$, define the operators $\underline{T}_\pi : \mathbb{R}_+^{|X|} \to \mathbb{R}_+^{|X|}$ and $\overline{T}_\pi : \mathbb{R}_+^{|X|} \to \mathbb{R}_+^{|X|}$ as follows: for $\underline{V}, \overline{V} \in \mathbb{R}_+^{|X|}$ and for $x \in X$,

$$\begin{aligned}&\underline{T}_\pi(\underline{V})(x) = \\ &\quad R(x,\pi(x)) + \gamma \min_{p \in p(x,\pi(x))} \sum_{y \in X} p(y|x,\pi(x))\underline{V}(y)\end{aligned} \quad (1)$$

and

$$\begin{aligned}&\overline{T}_\pi(\overline{V})(x) = \\ &\quad R(x,\pi(x)) + \gamma \max_{p \in p(x,\pi(x))} \sum_{y \in X} p(y|x,\pi(x))\overline{V}(y).\end{aligned} \quad (2)$$

It has been shown by Kurano *et al.* that for any $\pi \in \Pi$ with value function $V^\pi = [\underline{V}^\pi, \overline{V}^\pi]$, $\underline{V}^\pi$ and $\overline{V}^\pi$ are unique fixed points of $\underline{T}^\pi$ and $\overline{T}^\pi$, respectively; i.e.,

$$\begin{aligned}\underline{T}_\pi(\underline{V}^\pi)(x) &= \underline{V}^\pi(x), \quad x \in X \\ \overline{T}_\pi(\overline{V}^\pi)(x) &= \overline{V}^\pi(x), \quad x \in X.\end{aligned}$$

We remark that for each $\pi \in \Pi$, $\mathcal{P}(\pi) = \text{conv}\{\mathcal{P}^{(l)}(\pi) : l = 1, \ldots, l(\pi)\}$ for some $\mathcal{P}^{(l)}(\pi) \in \mathcal{P}(\pi), l = 1, \ldots, l(\pi)$, where for $\mathcal{D} \subset \mathbb{R}^{|X| \times |X|}$, conv($\mathcal{D}$) is the closed convex hull of $\mathcal{D}$ (see Lemma 1.1 in [5]). Thus, we can rewrite (1) as

$$\begin{aligned}&\underline{T}_\pi(\underline{V})(x) = \\ &\quad R(x,\pi(x)) + \gamma \min_{1 \le l \le l_{x,\pi(x)}} \sum_{y \in X} p^{(l)}(y|x,\pi(x))\underline{V}(y)\end{aligned}$$

for some $l_{x,\pi(x)}$, and similarly for the $\overline{T}_\pi$-operator. This leads to some computational simplification in applying the operators.

We present below two versions of our multi-policy improvement method for controlled Markov set-chains based on the idea of parallel rollout [1] for MDPs. The name "parallel rollout" comes from the idea that we "roll out" or simulate each available policy to estimate its value in parallel and then apply the most "promising" action to the system in on-line manner. The parallel-rollout method generalizes the policy-improvement step of the PI algorithm to the case of multiple policies.

Given a nonempty set $\Delta \subseteq \Pi$, we define $\Phi = [\underline{\Phi}, \overline{\Phi}] \in C(\mathbb{R}_+^{|X|})$ by

$$\begin{aligned}\underline{\Phi}(x) &= \max_{\pi \in \Delta} \underline{V}^\pi(x), \quad x \in X &(3)\\ \overline{\Phi}(x) &= \max_{\pi \in \Delta} \overline{V}^\pi(x), \quad x \in X. &(4)\end{aligned}$$

Next, for each $x \in X$, we define the two sets $\overline{A}_x$ and $\underline{A}_x$ as

$$\begin{aligned}\overline{A}_x = \arg\max_{a \in A} \Big\{ & R(x,a) \\ & + \gamma \max_{p \in p(x,a)} \sum_{y \in X} p(y|x,a)\overline{\Phi}(y) \Big\}\end{aligned} \quad (5)$$

and

$$\begin{aligned}\underline{A}_x = \arg\max_{a \in A} \Big\{ & R(x,a) \\ & + \gamma \min_{p \in p(x,a)} \sum_{y \in X} p(y|x,a)\underline{\Phi}(y) \Big\}.\end{aligned} \quad (6)$$

We then define a *parallel rollout policy* $\pi_{\text{pr}} \in \Pi$ with respect to Δ to be any policy such that for all $x \in X$,

$$\pi_{\text{pr}}(x) \in \begin{cases} \overline{A}_x \cap \underline{A}_x \text{ if } \overline{A}_x \cap \underline{A}_x \neq \emptyset \\ A \text{ otherwise }. \end{cases} \quad (7)$$

Define the *improvable state set* $\mathcal{I} = \{x | \overline{A}_x \cap \underline{A}_x \neq \emptyset, x \in X\}$. The following theorem establishes that over the set $\mathcal{I}$, the value of π_{pr} exceeds those of all policies in Δ. In other words, by following π_{pr}, *both the lower (value) bounds and the upper bounds of all policies* in Δ are improved simultaneously, provided the initial state belongs to the improvable state set.

Theorem 1: For a given nonempty set $\Delta \subseteq \Pi$, and for the policy π_{pr} defined in (7),

$$V^{\pi_{\text{pr}}}(x) \ge \max_{\pi \in \Delta} V^\pi(x), \quad x \in \mathcal{I},$$

where the max operator is defined componentwise.

To prove the above theorem, we begin with a lemma, which is similar to Corollaries 4.1 and 4.2 in [5].

Lemma 1: For any $V \in \mathbb{R}_+^{|X|}$ and $\pi \in \Pi$, if

$$\underline{T}_\pi(V)(x) \ge V(x), \quad x \in X \quad (8)$$

then $\underline{V}^\pi(x) \ge V(x)$, $x \in X$. Similarly, if

$$\overline{T}_\pi(V)(x) \ge V(x), \quad x \in X$$

then $\overline{V}^\pi(x) \ge V(x)$, $x \in X$.

Proof: By successive applications of the $\underline{T}_\pi$-operator to both sides of (8), and the monotonicity property of the operator (see Theorem 3.1 in [5]), we have that for all $x \in X$,

$$\lim_{n \to \infty} \underline{T}_\pi^n(V)(x) \ge V(x).$$

By Theorem 3.1 in [5], iterative application of $\underline{T}_\pi$ on any initial value function converges monotonically to the fixed point $\underline{V}^\pi$; i.e., $\lim_{n \to \infty} \underline{T}_\pi^n(V)(x) = \underline{V}^\pi(x)$, $x \in X$. The same argument applies to the $\overline{T}_\pi$ case and this proves the lemma. ■

We now prove the statement of Theorem 1.

Proof of Theorem 1: For any $x \in \mathcal{I}$ and any $\pi \in \Delta$, we have

$$\begin{aligned}&\underline{T}_{\pi_{\text{pr}}}(\underline{\Phi})(x) = \\ &\quad = R(x,\pi_{\text{pr}}(x)) \\ &\qquad\qquad + \gamma \min_{p \in p(x,\pi_{\text{pr}}(x))} \sum_{y \in X} p(y|x,\pi_{\text{pr}}(x))\underline{\Phi}(y) \\ &\quad = \max_{a \in A}\Big(R(x,a) + \gamma \min_{p \in p(x,a)} \sum_{y \in X} p(y|x,a)\underline{\Phi}(y)\Big) \\ &\qquad\qquad \text{by definition of } \pi_{\text{pr}}, \pi_{\text{pr}}(x) \in \overline{A}_x \cap \underline{A}_x \\ &\quad \ge R(x,\pi(x)) + \gamma \min_{p \in p(x,\pi(x))} \sum_{y \in X} p(y|x,\pi(x))\underline{\Phi}(y) \\ &\quad \ge R(x,\pi(x)) + \gamma \min_{p \in p(x,\pi(x))} \sum_{y \in X} p(y|x,\pi(x))\underline{V}^\pi(y)\end{aligned}$$

$$\begin{aligned} & \qquad \text{by definition of } \underline{\Phi} \\ = \;& \underline{V}^{\pi}(x). \end{aligned}$$

Therefore,

$$\underline{T}_{\pi_{\rm pr}}(\underline{\Phi})(x) \geq \underline{\Phi}(x), \quad x \in \mathcal{I}.$$

By Lemma 1, for any $x \in \mathcal{I}$ and any $\pi \in \Delta$,

$$\underline{V}^{\pi_{\rm pr}}(x) \geq \max_{\pi \in \Delta} \underline{V}^{\pi}(x).$$

Similarly, for any $x \in \mathcal{I}$ and any $\pi \in \Delta$,

$$\begin{aligned} & \overline{T}_{\pi_{\rm pr}}(\overline{\Phi})(x) = \\ & \quad R(x, \pi_{\rm pr}(x)) + \gamma \max_{p \in p(x, \pi_{\rm pr}(x))} \sum_{y \in X} p(y|x, \pi_{\rm pr}(x)) \overline{\Phi}(y) \\ \geq \;& R(x, \pi(x)) + \gamma \max_{p \in p(x, \pi(x))} \sum_{y \in X} p(y|x, \pi(x)) \overline{\Phi}(y) \\ & \qquad \text{by definition of } \pi_{\rm pr} \\ \geq \;& R(x, \pi(x)) + \gamma \max_{p \in p(x, \pi(x))} \sum_{y \in X} p(y|x, \pi(x)) \overline{V}^{\pi}(y) \\ & \qquad \text{by definition of } \overline{\Phi} \\ = \;& \overline{V}^{\pi}(x). \end{aligned}$$

Therefore, for any $x \in \mathcal{I}$ and any $\pi \in \Delta$,

$$\overline{V}^{\pi_{\rm pr}}(x) \geq \max_{\pi \in \Delta} \overline{V}^{\pi}(x). \tag{9}$$

The above implies that for all $\pi \in \Delta$ and $x \in \mathcal{I}$,

$$\begin{aligned} V^{\pi_{\rm pr}}(x) &= [\underline{V}^{\pi_{\rm pr}}(x), \overline{V}^{\pi_{\rm pr}}(x)], \\ &\geq [\underline{V}^{\pi}(x), \overline{V}^{\pi}(x)], \\ &= V^{\pi}(x). \end{aligned}$$

Therefore, for any $x \in \mathcal{I}$,

$$V^{\pi_{\rm pr}}(x) \geq \max_{\pi \in \Delta} V^{\pi}(x),$$

which completes the proof. ■

We remark that at state $x \in X - \mathcal{I}$, we have the freedom to choose any action. If we choose $\pi_{\rm pr}(x) \in \underline{A}_x$ for $x \in X - \mathcal{I}$, by following the parallel rollout policy, we can improve at least the lower bounds of all policies in Δ (or the upper bounds of all policies in Δ by choosing an action in $\overline{A}_x$). In that case, by the partial order, there does not exist $\pi \in \Delta$ such that $V^{\pi} > V^{\pi_{\rm pr}}$ if there exists at least one state $x \in X - \mathcal{I}$ such that $\underline{V}^{\pi_{\rm pr}}(x) > \max_{\pi \in \Delta} \underline{V}^{\pi}(x)$.

We now consider another policy defined from Δ for multi-policy improvement with the Φ-function defined by (3) and (4). Given $\Delta \subseteq \Pi$, define

$$\begin{aligned} \overline{A}_x[\Delta] &= \{\pi | \pi \in \Pi, \pi(x) \in \overline{A}_x, \forall x \in X\} \qquad (10) \\ \underline{A}_x[\Delta] &= \{\pi | \pi \in \Pi, \pi(x) \in \underline{A}_x, \forall x \in X\}, \qquad (11) \end{aligned}$$

where $\overline{A}_x$ and $\underline{A}_x$ are given in (5) and (6), respectively. Define $\tilde{\pi}_{\rm pr}$ to be any policy such that for $x \in X$,

$$\begin{aligned} \tilde{\pi}_{\rm pr}(x) \in \arg\max_{a \in \underline{A}_x} \Bigg\{ & R(x, a) \\ & + \gamma \max_{p \in p(x,a)} \sum_{y \in X} p(y|x, a) \max_{\pi \in \underline{A}_x[\Delta]} \overline{V}^{\pi}(y) \Bigg\}. \end{aligned} \tag{12}$$

The policy $\tilde{\pi}_{\rm pr}$ can be interpreted as follows. By defining $\tilde{\pi}_{\rm pr}(x)$, $x \in X$, over $\underline{A}_x$, it first improves all of the lower bounds of the policies in Δ. Then, $\tilde{\pi}_{\rm pr}$ does its best to improve the upper bounds of all of policies ϕ that improve all of the lower bounds of the policies in Δ:

$$\begin{aligned} \phi(x) \in \arg\max_{a \in A} \Bigg\{ & R(x, a) \\ & + \gamma \min_{p \in p(x,a)} \sum_{y \in X} p(y|x, a) \underline{\Phi}(y) \Bigg\}. \end{aligned} \tag{13}$$

The following theorem is a "localized" version of Theorem 4.1 in [5]. For a given nonempty $\Delta \subset \Pi$, we say that $\tilde{\pi}$ *is optimal with respect to* Δ if there does not exist $\pi \in \Delta$ such that $V^{\tilde{\pi}} < V^{\pi}$. (If $\Delta = \Pi$, then $\tilde{\pi}$ is an optimal policy for M in the usual "global" sense.)

Theorem 2: For any given nonempty set $\Delta \subseteq \Pi$ and for the policy $\tilde{\pi}_{\rm pr}$ defined in (12), $\tilde{\pi}_{\rm pr}$ is optimal with respect to Δ.

Proof: We first show that for all $\pi \in \Pi$ with $\pi(x) \in \underline{A}_x, x \in X$,

$$\overline{V}^{\tilde{\pi}_{\rm pr}}(x) \geq \overline{V}^{\pi}(x), \quad x \in X$$

and

$$\underline{V}^{\tilde{\pi}_{\rm pr}}(x) \geq \max_{\pi \in \Delta} \underline{V}^{\pi}(x), \quad x \in X.$$

(The argument is similar to that of the proof of Theorem 1. We only show the essential steps here.)

Define $\Psi = [\underline{\Psi}, \overline{\Psi}] \in C(\mathbb{R}_+^{|X|})$ such that $\underline{\Psi}(x) = \max_{\pi \in \Delta} \underline{V}^{\pi}(x)$, $x \in X$ and $\overline{\Psi}(x) = \max_{\pi \in \underline{A}_x[\Delta]} \overline{V}^{\pi}(x)$, $x \in X$. For any $x \in X$ and any $\pi \in \Delta$, we have

$$\begin{aligned} & \underline{T}_{\tilde{\pi}_{\rm pr}}(\underline{\Psi})(x) \\ = \;& R(x, \tilde{\pi}_{\rm pr}(x)) \\ & \quad + \gamma \min_{p \in p(x, \tilde{\pi}_{\rm pr}(x))} \sum_{y \in X} p(y|x, \tilde{\pi}_{\rm pr}(x)) \underline{\Psi}(y) \\ \geq \;& R(x, \pi(x)) + \gamma \min_{p \in p(x, \pi(x))} \sum_{y \in X} p(y|x, \pi(x)) \underline{\Psi}(y) \\ & \qquad \text{by definition of } \tilde{\pi}_{\rm pr}, \tilde{\pi}_{\rm pr}(x) \in \underline{A}_x \\ \geq \;& R(x, \pi(x)) + \gamma \min_{p \in p(x, \pi(x))} \sum_{y \in X} p(y|x, \pi(x)) \underline{V}^{\pi}(y) \\ = \;& \underline{V}^{\pi}(x). \end{aligned}$$

Therefore, by Lemma 1, for any $x \in X$ and any $\pi \in \Delta$,

$$\underline{V}^{\tilde{\pi}_{\rm pr}}(x) \geq \max_{\pi \in \Delta} \underline{V}^{\pi}(x).$$

Similarly, for any $x \in X$ and any $\pi \in \underline{A}_x[\Delta]$,

$$\begin{aligned} & \overline{T}_{\tilde{\pi}_{\rm pr}}(\overline{\Psi})(x) = \\ \geq \;& R(x, \pi(x)) + \gamma \max_{p \in p(x, \pi(x))} \sum_{y \in X} p(y|x, \pi(x)) \overline{\Psi}(y) \\ \geq \;& R(x, \pi(x)) + \gamma \max_{p \in p(x, \pi(x))} \sum_{y \in X} p(y|x, \pi(x)) \overline{V}^{\pi}(y) \\ = \;& \overline{V}^{\pi}(x), \end{aligned}$$

which implies that for all $\pi \in \Pi$ with $\pi(x) \in \underline{A}_x$,

$$\overline{V}^{\tilde{\pi}_{\rm pr}}(x) \geq \overline{V}^{\pi}(x), \pi(x) \in \underline{A}_x, \quad x \in X. \qquad (14)$$

Now observe that if there exists $\pi \in \Delta$ such that $V^{\tilde{\pi}_{\rm pr}} < V^{\pi}$, then the following must be true: $\underline{V}^{\tilde{\pi}_{\rm pr}}(x) \leq \underline{V}^{\pi}(x), x \in X$. But this is impossible unless $\underline{V}^{\pi} = \underline{V}^{\tilde{\pi}_{\rm pr}}$. If that is the case, then $\pi(x) \in \underline{A}_x$, $x \in X$. To see this,

$$\begin{aligned}
&\underline{V}^{\pi}(x) = \underline{V}^{\tilde{\pi}_{\rm pr}}(x) \\
&\quad = \quad R(x, \tilde{\pi}_{\rm pr}(x)) \\
&\qquad\qquad + \gamma \min_{p \in p(x, \tilde{\pi}_{\rm pr}(x))} \sum_{y \in X} p(y|x, \tilde{\pi}_{\rm pr}(x)) \underline{V}^{\tilde{\pi}_{\rm pr}}(y) \\
&\quad = \quad R(x, \tilde{\pi}_{\rm pr}(x)) \\
&\qquad\qquad + \gamma \min_{p \in p(x, \tilde{\pi}_{\rm pr}(x))} \sum_{y \in X} p(y|x, \tilde{\pi}_{\rm pr}(x)) \underline{V}^{\pi}(y) \\
&\quad = \quad R(x, \tilde{\pi}_{\rm pr}(x)) \\
&\qquad\qquad + \gamma \min_{p \in p(x, \tilde{\pi}_{\rm pr}(x))} \sum_{y \in X} p(y|x, \tilde{\pi}_{\rm pr}(x)) \max_{\pi' \in \Delta} \underline{V}^{\pi'}(y) \\
&\quad = \quad \max_{a \in A} \Bigg(R(x, a) \\
&\qquad\qquad + \gamma \min_{p \in p(x,a)} \sum_{y \in X} p(y|x, a) \max_{\pi' \in \Delta} \underline{V}^{\pi'}(y) \Bigg) \\
&\quad = \quad R(x, \pi(x)) + \gamma \min_{p \in p(x, \pi(x))} \sum_{y \in X} p(y|x, \pi(x)) \underline{V}^{\pi}(y),
\end{aligned}$$

which implies that $\pi(x) \in \underline{A}_x$.

By (14), there is no $x \in X$ such that $\overline{V}^{\tilde{\pi}_{\rm pr}}(x) < \overline{V}^{\pi}(x)$ for any $\pi \in \underline{A}_x[\Delta]$. Therefore, there does not exist $\pi \in \Delta$ such that $V^{\tilde{\pi}_{\rm pr}} < V^{\pi}$. This implies that $\tilde{\pi}_{\rm pr}$ is optimal with respect to Δ, which completes the proof. ■

We remark that the following policy symmetrically defined as

$$\begin{aligned}
\tilde{\pi}_{\rm pr}(x) \in \arg\max_{a \in \overline{A}_x} \Bigg\{ & R(x, a) \\
& + \gamma \min_{p \in p(x,a)} \sum_{y \in X} p(y|x, a) \max_{\pi \in \overline{A}_x[\Delta]} \overline{V}^{\pi}(y) \Bigg\} \quad (15)
\end{aligned}$$

for $x \in X$ is also optimal with respect to Δ.

Based on the above policy improvement results, the following PI-type algorithm follows naturally. (Below, we use (15); alternatively, the algorithm can be constructed using (12) instead of (15)).

0. **Initialization:** $k = 0$ and set π_0 arbitrarily in Π.
1. **Policy Evaluation:** obtain V^{π_k}.
2. **Policy Improvement:** obtain π_{k+1} by (15) with $\Delta = \{\pi_k\}$.
3. **Stop Condition:** If $\overline{V}^{\pi_k} = \overline{V}^{\pi_{k+1}}$, then stop. Otherwise, $k \leftarrow k + 1$ and go to step 1.

Several observations can be made on this PI-type algorithm. First, the upper bounds are monotonically increasing, i.e., for all k,

$$\overline{V}^{\pi^k}(x) \leq \overline{V}^{\pi^{k+1}}(x), \quad x \in X.$$

Therefore, π_k cannot beat π_{k+1} in terms of the partial order. In fact, π_k *is optimal with respect to* $\{\pi_0, \pi_1, \ldots, \pi_{k-1}\}$. Because there are finitely many policies and the upper bounds are monotonically increasing, the algorithm converges to a policy in a finite number of steps—$|\Pi|$ steps in the worst case. Let $\overline{\pi}^*$ be the converged policy. We know that $\overline{\pi}^*$ achieves the best upper bound; i.e., for all $\pi \in \Pi$, $\overline{V}^{\overline{\pi}^*}(x) \geq \overline{V}^{\pi}(x)$, $x \in X$.

Pick any optimal policy $\pi^* \in \Pi$. It cannot happen that there exists a state $x \in X$ such that $\underline{V}^{\overline{\pi}^*}(x) > \underline{V}^{\pi^*}(x)$ because if that were the case, it would contradict with the definition of optimality. Therefore,

$$\underline{V}^{\overline{\pi}^*}(x) \leq \underline{V}^{\pi^*}(x), \quad x \in X.$$

From (15), $\overline{\pi}^*$ improves the lower bounds of all of the policies (including $\overline{\pi}^*$) that improve the upper bound of the best-upper-bound-achieving policy $\overline{\pi}^*$. Suppose that there exists an optimal policy that achieves the best upper-bound; denote it as $\overline{\rho}$. Then,

$$\underline{V}^{\overline{\pi}^*}(x) \geq \underline{V}^{\overline{\rho}}(x), \quad x \in X.$$

These observations imply that $\overline{\pi}^*$ achieves $\underline{V}^{\overline{\rho}}$; i.e., $\underline{V}^{\overline{\pi}^*} = \underline{V}^{\overline{\rho}}$. Therefore, it must be true that $\overline{V}^{\overline{\pi}^*} = \overline{V}^{\overline{\rho}}$. Otherwise, $\overline{\rho}$ cannot be an optimal policy. Hence, we have that $V^{\overline{\rho}} = V^{\overline{\pi}^*}$.

We can apply a similar reasoning to the PI-type algorithm with (12), but with the following **Stop Condition**: If $\underline{V}^{\pi_k} = \underline{V}^{\pi_{k+1}}$, then stop. Otherwise, $k \leftarrow k + 1$ and go to step 1.

Summarizing, *if there exists an optimal policy* $\overline{\rho}$ *(or* $\underline{\rho}$*) in the partial order that achieves the best upper-bound function (or, resp., the best lower-bound function), then the PI-type algorithm with (15) (resp. (12)) converges to a policy* $\overline{\pi}^*$ *(resp.* $\underline{\pi}^*$*) such that* $V^{\overline{\rho}} = V^{\overline{\pi}^*}$ *(resp.* $V^{\underline{\rho}} = V^{\underline{\pi}^*}$*) with the corresponding stop condition.*

We can extend the above single-policy improvement algorithm into a multi-policy improvement algorithm. At iteration $k \geq 0$, we have a set of policies $\Pi_k \in \Pi$, instead of a single policy, and we evaluate each policy $\pi_k \in \Pi_k$ and obtain π_{k+1} by applying (12) with $\Delta = \Pi_k$. We then set $\Pi_{k+1} = \{\pi_{k+1}\} \cup \Omega_{k+1}$, where Ω_{k+1} is an arbitrary subset of Π. Note that we have significant freedom in our choice of Ω_k. One possibility is to run the single-policy improvement method in parallel and put the sequence of the policies generated into Ω_k.

The result of Theorem 1 also gives rise to a (heuristic) PI-type algorithm if we use (7) instead of (12) or (15), with the following stop condition: if either $\underline{V}^{\pi_k} = \underline{V}^{\pi_{k+1}}$ or $\overline{V}^{\pi_k} = \overline{V}^{\pi_{k+1}}$, then stop (alternative conditions are also possible). In this case, π_{k+1} improves both the lower and upper bounds of π_k over the improvable state set of π_k. Therefore, a converged policy, if it exists, improves all of the policies that have appeared in the algorithm, over the intersection of their improvable state sets.

IV. Applications

Even though the two well-known algorithms—value iteration and policy iteration—are available for solving MDPs, it is generally understood that solving MDPs with large state and/or action spaces in practice using these algorithms is impossible. Numerous efforts have been devoted to solving MDPs approximately and/or heuristically. Similarly, we can expect that solving controlled Markov set-chains with large state and/or action spaces via the value-iteration type algorithm studied by Kurano *et al.* is also practically difficult.

Following the spirit of parallel rollout [1], it is often true that for a given problem, we already have some heuristic policies available. 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. The results presented in the previous section are directly relevant to this goal.

Suppose we have a large MDP $M_o = (X, A, P, R)$, where P is the state transition matrix. We then partition the state space X into $B = \{B_1, \ldots, B_n\}$ with nonempty $B_i \subset X$, $\bigcup_{i=1}^{n} B_i = X$, and $B_i \cap B_j = \emptyset$ for $i \neq j, i, j = 1, \ldots, n$. We then construct a controlled Markov set-chain model $M_r = (B, A, P_M, R_M)$ such that the state space is B, the action space is A, and $P_M = \langle \underline{p}_M, \overline{p}_M \rangle$ is defined such that for $B_i, B_j \in B$ and $a \in A$,

$$\begin{aligned} \underline{p}_M(B_j|B_i, a) &= \min_{x \in B_i} \sum_{y \in B_j} P(y|x, a) \\ \overline{p}_M(B_j|B_i, a) &= \max_{x \in B_i} \sum_{y \in B_j} P(y|x, a). \end{aligned}$$

We can then consider a "minimum" (pessimistic) model with $R_M(B_i, a) = \min_{x \in B_i} R(x, a)$ or a "maximum" (optimistic) model with $R_M(B_i, a) = \max_{x \in B_i} R(x, a)$. Given a set of heuristic policies $\{\pi | \pi : B \rightarrow A\}$, we can then apply our multi-policy improvement method to the model M_r. Note that the V^{π}-function for a given policy π for M_r provides a performance bound for the corresponding policy applied to M_o. (If a heuristic policy is given for M_r as a mapping from B to A, to apply it to M_o, we induce a corresponding policy as a mapping from X to A.) By applying the multi-policy improvement method, we can improve the performance bounds of the available policies.

To use the multi-policy improvement method for on-line control of a Markov set-chain (following the method of parallel rollout), we need to estimate $V^{\pi}(x)$ for the current state $x \in X$ and $\pi \in \Pi$ via simulation. Unlike in the MDP case, a direct application of Monte Carlo simulation would not apply here. Developing an efficient simulation method to estimate $V^{\pi}(x)$ is an interesting research topic. Here we discuss one heuristic method.

We simulate π as follows: at state x we take $\pi(x)$ and select l from $1 \leq l \leq l_{x,\pi(x)}$ uniformly, and make a transition to the next state according to $p^{(l)}(\cdot|x, \pi(x))$, thereby generating a single simulated sample path over a finite horizon. We repeat this over many sample paths and obtain the estimated interval of following the policy π with an initial state x.

For on-line control, we evaluate the value of taking each action $a \in A$ at state x_t at time t as follows. We sample a set of next states $y_1^a, \ldots, y_n^a$ randomly, uniformly over X, and for each $\pi \in \Delta$, we estimate the value of following policy π (which we denote $\hat{V}^{\pi}(y_i^a)$). Then, we select at time t an action a in

$$\arg\max_{a \in \underline{\hat{A}}_x} \Bigg\{ R(x_t, a) + \frac{\gamma}{n} \times \max_{p \in p(x_t,a)} \sum_{i=1}^{n} p(y_i^a|x, a) \max_{\pi \in \hat{\Delta}} \overline{\hat{V}}^{\pi}(y_i^a) \Bigg\},$$

where $\hat{\Delta} = \{\pi | \pi \in \Pi, \pi(x) \in \underline{\hat{A}}_x, \forall x \in X\}$ and

$$\underline{\hat{A}}_x = \arg\max_{a \in A} \Bigg\{ R(x_t, a) + \frac{\gamma}{n} \times \min_{p \in p(x_t,a)} \sum_{i=1}^{n} p(y_i^a|x_t, a) \max_{\pi \in \Delta} \underline{\hat{V}}^{\pi}(y_i^a) \Bigg\}.$$

V. Concluding Remarks

"Policy switching" [1] is yet another multi-policy improvement method for MDPs, where the maximization over the action space is not necessary (cf. parallel rollout), making the method attractive for problems with large action spaces. A future direction is to extend the original policy switching method to the case of controlling Markov set-chains.

An issue that remains unanswered is a characterization of the performance gains to be expected by using the proposed method. Providing analytical results along these lines is challenging. Indeed, to the best of our knowledge, no analytical result exists on performance gains for single policy improvement in PI for MDPs.

References

[1] 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, pp. 309–341, 2004.

[2] R. Givan, S. Leach, and T. Dean, "Bounded Markov decision processes," *Artificial Intelligence*, vol. 122, pp. 71–109, 2000.

[3] D. J. Hartfiel, *Markov Set-Chains*, Springer, Berlin, 1998.

[4] S. Kalyanasundaram, E. K. P. Chong, and N. B. Shroff, "Markov decision processes with uncertain transition rates: Sensitivity and max-min control," *Asian Journal of Control, special issue on Control of Discrete Event Systems,* vol. 6, no. 2, pp. 253–269, 2004.

[5] M. Kurano, J. Song, M. Hosaka, and Y. Huang, "Controlled Markov set-chains with discounting," *J. Appl. Prob.*, vol. 35, pp. 293–302, 1998.

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

[7] J. K. Satia and R. E. Lave, "Markovian decision processes with uncertain transition probabilities," *Operations Research*, vol. 21, pp. 728–740, 1973.

[8] C. C. White and H. K. Eldeib, "Markov decision processes with imprecise transition probabilities," *Operations Research*, vol. 43, pp. 739–749, 1994.