Automatica 49 (2013) 3687–3689

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Technical communique

Policy set iteration for Markov decision processes☆

Hyeong Soo Chang [1]

Department of Computer Science and Engineering, Sogang University, Seoul, Republic of Korea

ARTICLE INFO

Article history:
Received 6 November 2012
Received in revised form
20 April 2013
Accepted 3 September 2013
Available online 4 October 2013

Keywords:
Markov decision processes
Policy iteration
Dynamic programming
Randomization

ABSTRACT

This communique presents an algorithm called "policy set iteration" (PSI) for solving infinite horizon discounted Markov decision processes with finite state and action spaces as a simple generalization of policy iteration (PI). PSI generates a monotonically improving sequence of stationary Markovian policies $\{\pi_k^*\}$ based on a set manipulation, as opposed to PI's single policy manipulation, at each iteration k. When the set involved with PSI at k contains N independently generated sample-policies from a given distribution d, the probability that the expected value of any sampled policy from d with respect to an initial state distribution is greater than that of π_k^* converges to zero with $O(N^{-k})$ rate. Moreover, PSI converges to an optimal policy no slower than PI in terms of the number of iterations for any d.

1. Introduction

Consider a Markov decision process (MDP) (Puterman, 1994) $M = (X, A, P, R)$, where X is a finite state set, $A(x)$ is a finite action set at $x \in X$ with $\bigcup_{x\in X} A(x) = A$, R is a reward function such that $R(x, a) \in \mathbb{R}$, $x \in X$, $a \in A(x)$, and P is a transition function that maps $\{(x, a)|x \in X, a \in A(x)\}$ to the set of probability distributions over X. We denote the probability of making a transition to state $y \in X$ when taking an action $a \in A(x)$ at state $x \in X$ by P_{xy}^a.

We define a (stationary Markovian) policy π as a mapping from X to A with $\pi(x) \in A(x)$, $\forall x \in X$, and let Π be the set of all such policies. Define *the value of* $\pi \in \Pi$ with an initial state $x \in X$:

$$V^\pi(x) = E\left[\sum_{t=0}^{\infty} \gamma^t R(X_t, \pi(X_t)) \middle| X_0 = x\right],$$

where X_t is a random variable denoting state at time t by following π and $\gamma \in (0, 1)$ is a discounting factor.

Our goal is to find an optimal policy $\pi^* \in \Pi$ that achieves the optimal value $V^*(x) = \max_{\pi\in\Pi} V^\pi(x)$ at all $x \in X$.

Policy iteration (PI) (Howard, 1960; Puterman, 1994) is the well-known exact search algorithm in policy space for solving this problem. Each iteration consists of two parts: policy evaluation and policy improvement. The policy evaluation step obtains V^π for a given $\pi \in \Pi$ and the policy improvement step takes V^π and obtains a new policy π' such that

$$\pi'(x) \in \arg\max_{a\in A(x)} \left(R(x, a) + \gamma \sum_{y\in X} P_{xy}^a V^\pi(y)\right), \quad x \in X,$$

and this step ensures monotonicity in terms of the policy performance, i.e., $V^{\pi'}(x) \geq V^\pi(x)$ for all $x \in X$. Thus PI's single iteration has $O(|X|^2|A| + |X|^3)$ time-complexity. Starting with an arbitrary $\pi_0 \in \Pi$, at each iteration $k \geq 1$, PI applies the policy evaluation and policy improvement steps alternately until $V^{\pi_k}(x) = V^{\pi_{k-1}}(x)$ for all $x \in X$, in which case an optimal policy has been found. Because Π is finite, PI guarantees convergence to an optimal solution in a finite number of iterations.

Even if it is known that PI often outperforms value iteration (VI) in practical applications (Blondel & Tsitsiklis, 2000) (in particular, considerably better than VI for small-scale problems (state space size less than 10,000), provided that γ is close to 1 (Rust, 1994)), (Hollanders, Delvenne, & Jungers, 2012) has recently shown that the worst-case complexity of PI for infinite horizon discounted MDPs is exponential in general. That is, there exist MDPs in which PI requires an exponential number of iterations to converge. Furthermore, to the author's best knowledge, little is known about the convergence rate of PI in terms of the number of iterations except that $\{V^{\pi_k}\}$ for $\{\pi_k\}$ generated by PI converges to V^* monotonically, i.e., with a linear rate, even though there exists sufficient conditions for a quadratic rate (Puterman, 1994, p. 181).

In this communique, we present a novel algorithm called "policy set iteration" (PSI) for solving M as a simple generalization of PI. PSI generates a sequence of policies $\{\pi_k^*\}$ based on *a set manipulation*, as opposed to PI's single policy manipulation, at each iteration k. Similar to PI, PSI converges to an optimal policy in a finite

☆ The material in this paper was not presented at any conference. This paper was recommended for publication in revised form by Associate Editor Nuno C. Martins under the direction of Editor André L. Tits.

E-mail address: hschang@sogang.ac.kr.

[1] Tel.: +82 2 705 8925; fax: +82 2 704 8273.

0005-1098/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.automatica.2013.09.010

number of iterations. In particular, when the set involved with PSI at k contains $N \geq 0$ independently generated sample-policies from a given distribution d, the probability that V_δ^π of any sampled policy π from d is greater than $V_\delta^{\pi_k^*}$ converges to zero with $O(N^{-k})$ rate, where $V_\delta^\pi = \sum_{x\in X} \delta(x)V^\pi(x)$, $\pi \in \Pi$, for a given initial state distribution δ over X. Moreover, PSI converges no slower than PI in terms of the number of iterations for any d if π_0^* is set to be the same initial policy chosen for PI.

Each iteration of PSI requires $O((N + m + 1)(|X|^2|A| + |X|^3))$ time-complexity if the set involved with PSI at each iteration contains N sample-policies and $m \geq 0$ additional arbitrarily chosen policies in Π. Therefore, we establish that by allowing an increment in the time-complexity of PI by a factor of about N, *no slower convergence than* PI can be achieved in terms of the number of iterations and *a finite-time probabilistic error-bound* in obtaining an optimal policy for a given δ can be derived. The key ideas are improving a set of multiple policies by using the multi-policy improvement method (Chang, Fu, Hu, & Marcus, 2013, Chapter 5) and incorporating randomization into the deterministic PI by including randomly generated sample-policies into the set for the improvement.

In fact, the approach of using probability and randomization for control of systems with uncertainty has been studied over the years in the area of research known as "probabilistic robust control" (see, e.g., a survey of Calafiore, Dabbene, and Tempo (2011)) to overcome computational problems in deterministic methods. Assuming that a probability measure of the uncertainty is given, the objective of this approach is to provide probabilistic assessments on the system characteristics, i.e., a certain performance level satisfied by the system in terms of probability. In particular, Calafiore and Campi developed the so-called scenario design method (Calafiore & Campi, 2006) to effectively solve control design problems that can be cast in the form of a convex optimization problem with uncertain constraints and they provided a probability bound of sample-maximum estimate of a random variable (Calafiore, 2009; Campi & Calafiore, 2009) as a result of applying the method. PSI takes the spirit of randomized methods in probabilistic robust control and the result of the probability bound of sample-maximum estimate is used in deriving the probabilistic convergence rate result of PSI.

Some variants of PI to expedite the computation speed of PI exist, e.g., Bertsekas (2013), Mrkaic (2002), Puterman and Shin (1978). These methods correspond to substituting exact policy evaluation with some successive approximation iteration, not concerned with improving the convergence rate in terms of iteration. Even though the exposition is done here with the assumption that V^π is exactly obtained for $\pi \in \Pi$ in policy evaluation of PSI, these (approximate) policy evaluation approaches can be incorporated into PSI.

2. Policy set iteration

We formally describe PSI below. If $N = 0$ and $\Psi_k = \emptyset$ for all $k \geq 1$, then PSI degenerates to PI. The structure of the algorithm basically follows that of PI. It consists of three main steps: policy-set evaluation, policy-set improvement, and policy-set generation.

Policy Set Iteration

1. **Initialization:** Select $N \geq 0$ and an arbitrary distribution d over Π. Obtain Δ_0 which contains N policies sampled from d and an arbitrary policy $\pi_0^* \in \Pi$. Set $k = 0$.
2. **Loop:**
   2.1 **Policy-set evaluation:** Obtain V^π for all $\pi \in \Delta_k$.
   2.2 **Policy-set improvement:** Obtain π_{k+1}^* such that for $x \in X$,
   $$\pi_{k+1}^*(x) \in \arg\max_{a\in A(x)} \left(R(x,a) + \gamma \sum_{y\in X} P_{xy}^a \max_{\pi\in\Delta_k} V^\pi(y) \right).$$
   2.3 **Policy-set generation:** $\Delta_{k+1} \leftarrow \{\pi_{k+1}^*\} \cup \Psi_k$ where Ψ_k contains N policies sampled from a distribution d over Π and possibly additional policies chosen in Π.
   2.4 If $V^{\pi_{k+1}^*}(x) = V^{\pi_k^*}(x)$ for all $x \in X$, exit the loop. Otherwise, $k \leftarrow k + 1$.

For the analysis of PSI, we start with showing that for $k \geq 0$, π_{k+1}^* improves all policies in $\Delta_0 \cup \Delta_1 \cup \cdots \cup \Delta_k$.

Theorem 1. *For $k \geq 0$,*

$$V^{\pi_{k+1}^*}(x) \geq \max_{\pi\in\bigcup_{l=0}^k \Delta_l} V^\pi(x)$$

for all $x \in X$ for any $\{\Psi_k\}$.

Proof. We use induction on k. For $k = 0$, by applying Theorem 5.12 of the multi-policy improvement in Chang et al. (2013), $V^{\pi_1^*}(x) \geq \max_{\pi\in\Delta_0} V^\pi(x)$ for all $x \in X$. Assume that at $k = m$, $V^{\pi_{m+1}^*}(x) \geq \max_{\pi\in\bigcup_{l=0}^m \Delta_l} V^\pi(x)$ for all $x \in X$. Then because $\Delta_{m+1} \leftarrow \{\pi_{m+1}^*\} \cup \Psi_m$, $V^{\pi_{m+2}^*}(x) \geq \max_{\pi\in\Delta_{m+1}} V^\pi(x) \geq V^{\pi_{m+1}^*}(x)$ for all $x \in X$. Therefore, $V^{\pi_{m+2}^*}(x) \geq \max_{\pi\in\bigcup_{l=0}^{m+1} \Delta_l} V^\pi(x)$ for all $x \in X$. And this argument holds for any $\{\Psi_k\}$.

Theorem 2. *PSI converges to an optimal policy in a finite number of iterations for any sequence of $\{\Psi_k, k \geq 0\}$.*

Proof. We use the classical reasoning in the convergence proof of PI.

First, by the previous result, we have that $V^{\pi_{k+1}^*}(x) \geq V^{\pi_k^*}(x)$ for all $x \in X$ for any sequence of $\{\Psi_k, k \geq 0\}$. Because $\{V^{\pi_k^*}\}$ are monotonically improving and Π is finite, PSI must terminate under the stopping condition. At the termination of $k = m$, $V^{\pi_{m+1}^*} = V^{\pi_m^*}$ and $V^{\pi_m^*}$ satisfies the optimality equation of

$$V^{\pi_m^*}(x) = \max_{a\in A(x)} \left(R(x,a) + \gamma \sum_{y\in X} P_{xy}^a V^{\pi_m^*}(y) \right)$$

for all $x \in X$, which implies that $V^{\pi_m^*} = V^*$ from the uniqueness of V^*. Therefore, π_m^* is an optimal policy.

The *worst-case* complexity of PSI is therefore equal to that of PI. It converges to an optimal policy in $|\Pi|$ iterations at the worst case. However, in PSI $\{\Psi_k\}$ provides considerable freedom of including arbitrary policies and π_{k+1}^* improves all policies in $\Delta_0 \cup \Delta_1 \cup \cdots \cup \Delta_k$, possibly making the convergence rate of PSI much faster than PI. In particular, by making Ψ_k include the policy generated by PI at iteration k and PI use the same initial policy as PSI's, PSI converges no slower than PI. The following result is immediate from the above theorem.

Corollary 3. *Let $\{\pi_k^{PI}, k \geq 0\}$ be the sequence of policies generated by running PI. If $\pi_k^{PI} \in \Psi_k$ for all $k \geq 1$ and $\pi_0^{PI} = \pi_0^*$, then PSI converges to an optimal policy no slower than PI in terms of the number of iterations.*

We now provide a finite-time probabilistic convergence rate of PSI when Ψ_k contains $N \geq 1$ independently generated sample-policies from a given distribution d over Π. Specifically, we show that the probability that the expected value of any sampled policy from d with respect to an initial state distribution is greater than that of π_k^* converges to zero with $O(N^{-k})$ rate. Recall that $V_\delta^\pi = \sum_{x\in X} \delta(x)V^\pi(x)$, $\pi \in \Pi$, for a given initial state distribution δ over X.

Theorem 4. *In PSI, suppose that $N \geq 1$. Consider a random variable Z defined on Π whose probability distribution is given by d. Then for*

any given initial state distribution δ over X and for any d,

$$\Pr\{V_\delta^Z > V_\delta^{\pi_k^*}\} \le \left(\frac{1}{N+1}\right)^k, \quad k \ge 1.$$

We stress that the measure Pr in the statement of the above theorem is the measure induced by the distribution d. Different choices of d can lead to different probabilities. However, the bound holds regardless of the choices of d.

Proof. Let $\{\psi_1^k, \ldots, \psi_N^k\}$ be the set of N independent sample-policies generated from d at k. We write Ψ_k as the union of $\{\psi_1^k, \ldots, \psi_N^k\}$ and α_k where α_k is the set of all policies that were not obtained by sampling from d at k. Fix any δ.

By Theorem 1, for $k \ge 0$, $V_\delta^{\pi_{k+1}^*} \ge \max_{\pi\in\Delta_k} V_\delta^\pi$ for δ. It follows that for any $k \ge 0$, the event $\{V_\delta^Z > V_\delta^{\pi_{k+1}^*}\}$ implies $\{V_\delta^Z > \max_{\{\pi_k^*,\psi_1^k,\ldots,\psi_N^k\}\cup\alpha_k} V_\delta^\pi\}$. Therefore, we have that

$$\begin{aligned}
\Pr\{V_\delta^Z > V_\delta^{\pi_{k+1}^*}\} &\le \Pr\left\{V_\delta^Z > \max_{\{\pi_k^*,\psi_1^k,\ldots,\psi_N^k\}\cup\alpha_k} V_\delta^\pi\right\} \\
&\le \Pr\{V_\delta^Z > V_\delta^{\pi_k^*}\} \times \Pr\left\{V_\delta^Z > \max_{\pi\in\alpha_k} V_\delta^\pi\right\} \\
&\quad \times \Pr\left\{V_\delta^Z > \max_{\{\psi_1^k,\ldots,\psi_N^k\}} V_\delta^\pi\right\} \\
&\le \Pr\{V_\delta^Z > V_\delta^{\pi_k^*}\} \times \Pr\left\{V_\delta^Z > \max_{\{\psi_1^k,\ldots,\psi_N^k\}} V_\delta^\pi\right\}.
\end{aligned}$$

By Campi and Calafiore (2009, Proposition 4), for a discrete random variable Y on a finite nonempty set Ω with any probability distribution d over Ω and $f : \Omega \to \Re$,

$$\Pr\left\{f(Y) > \max_{i=1,\ldots,N} f(y_i)\right\} \le \frac{1}{N+1}$$

where y_i's are N independent, identically distributed samples of Y from d. Therefore for any d, $\Pr\{V_\delta^Z > \max_{\{\psi_1^k,\ldots,\psi_N^k\}} V_\delta^\pi\} \le (N+1)^{-1}$ so that for $k \ge 0$ and any d,

$$\Pr\{V_\delta^Z > V_\delta^{\pi_{k+1}^*}\} \le \frac{1}{N+1} \times \Pr\{V_\delta^Z > V_\delta^{\pi_k^*}\}.$$

This finally implies that for $k \ge 1$ and any d,

$$\Pr\{V_\delta^Z > V_\delta^{\pi_k^*}\} \le \left(\frac{1}{N+1}\right)^k.$$

3. Concluding remark

PSI can be viewed as a *general framework* for combining multiple solution methods which generate a sequence of policies or a sequence of sets of policies and for converting any non-convergent heuristic method into a convergent algorithm. Those policies in the sequences generated by any other methods can be included into $\{\Psi_k\}$. By Theorems 1 and 2 then, π_{k+1}^* of the resulting PSI algorithm improves all policies in $\Delta_0 \cup \Delta_1 \cup \cdots \cup \Delta_k$ and converges to an optimal policy.

With a higher computational complexity per step, PSI has a possibly faster convergence rate than PI in terms of the number of iteration steps. The present communique advocates this by the theoretical analysis of a finite-time probabilistic error-bound. Quantifying the trade-off along with an experimental study of PSI is a good future work.

Because no structural assumptions can be made on the policy space, we established a general probability bound that holds for any choice of d. In practice, the simplest choice of d would be the uniform distribution or heuristically designed one if some structure of the policy space is available. Studying on how to set d in PSI is a good future research topic. There are various "model-based" optimization algorithms that generate a sequence of the probability distributions over the set of the solutions of a given problem, i.e., "learn" the solution space (see, e.g., the references in Hu, Hu, and Chang (2012)). It would be an interesting work to incorporate (intermediate) results of those algorithms into PSI.

References

Bertsekas, D. P. (2013). Lambda-policy iteration: a review and a new implementation. In F. Lewis, & D. Liu (Eds.), *Reinforcement learning and approximate dynamic programming for feedback control*. IEEE Press Computational Intelligence Series.

Blondel, V. D., & Tsitsiklis, J. N. (2000). A survey of computational complexity results in systems and control. *Automatica*, *36*, 1249–1274.

Calafiore, G. C. (2009). On the expected probability of constraint violation in sampled convex programs. *Journal of Optimization Theory and Applications*, *143*, 405–412.

Calafiore, G. C., & Campi, M. C. (2006). The scenario approach to robust control design. *IEEE Transactions on Automatic Control*, *51*, 742–753.

Calafiore, G. C., Dabbene, F., & Tempo, R. (2011). Research on probabilistic methods for control system design. *Automatica*, *47*, 1279–1293.

Campi, M. C., & Calafiore, G. C. (2009). Notes on the scenario design approach. *IEEE Transactions on Automatic Control*, *54*, 382–385.

Chang, H. S., Fu, M. C., Hu, J., & Marcus, S. I. (2013). *Simulation-based algorithms for Markov decision processes* (2nd ed.). London: Springer.

Hollanders, R., Delvenne, J-C., & Jungers, R. M. (2012). The complexity of policy iteration is exponential for discounted Markov decision processes. *Proc. of the 51st IEEE Conf. on decision and control*, 5997–6002.

Howard, R. A. (1960). *Dynamic programming and Markov processes*. Cambridge, MA: The MIT Press.

Hu, J., Hu, P., & Chang, H. S. (2012). A stochastic approximation framework for a class of randomized optimization algorithms. *IEEE Transactions on Automatic Control*, *57*, 165–178.

Mrkaic, M. (2002). Policy iteration accelerated with Krylov methods. *Journal of Economic Dynamics & Control*, *26*, 517–545.

Puterman, M. L. (1994). *Markov decision processes: discrete stochastic dynamic programming*. New York: Wiley.

Puterman, M. L., & Shin, M. C. (1978). Modified policy iteration algorithms for discounted Markov decision problems. *Management Science*, *24*, 1127–1137.

Rust, J. (1994). Structural estimation of Markov decision processes. In R. Engle, & D. McFadden (Eds.), *Handbook of econometrics*. Amsterdam: North-Holland/Elsevier.