Automatica 50 (2014) 1940–1943

Contents lists available at ScienceDirect

Automatica

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

Technical communique

Value 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 8 November 2013
Received in revised form
21 February 2014
Accepted 30 April 2014
Available online 4 June 2014

Keywords:
Markov decision processes
Value iteration
Dynamic programming
Constrained optimization

ABSTRACT

This communique presents an algorithm called "value set iteration" (VSI) for solving infinite horizon discounted Markov decision processes with finite state and action spaces as a simple generalization of value iteration (VI) and as a counterpart to Chang's policy set iteration. A sequence of value functions is generated by VSI based on manipulating a set of value functions at each iteration and it converges to the optimal value function. VSI preserves convergence properties of VI while converging no slower than VI and in particular, if the set used in VSI contains the value functions of independently generated sample-policies from a given distribution and a properly defined policy switching policy, a probabilistic exponential convergence rate of VSI can be established. Because the set used in VSI can contain the value functions of any policies generated by other existing algorithms, VSI is also a general framework of combining multiple solution methods.

1. Introduction

Consider a Markov decision process (MDP) (Puterman, 1994) (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^a_{xy}.

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 function* V^π of $\pi \in \Pi$ over X such that

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

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

Let $B(X)$ be the set of all real-valued functions on X and define a mapping $L : B(X) \to B(X)$ such that for $x \in X$ and $u \in B(X)$,

$$L(u)(x) := \max_{a\in A(x)} \left(R(x, a) + \gamma \sum_{y\in X} P^a_{xy} u(y)\right).$$

It is well-known that V^* uniquely satisfies $L(V^*) = V^*$ and the sequence of value functions $\{U_k\}$ generated by iterative applications of L with an arbitrary initial value function $U_0 \in B(X)$ such that $L(U_k) = U_{k+1}, k \geq 0$, converges to V^* and this exact method is called value iteration (VI) (Puterman, 1994). Because L is a contraction mapping in $B(X)$ with γ-contraction, VI's convergence rate is linear with the rate of γ and VI produces an ϵ-optimal policy for any given $\epsilon > 0$. The ϵ-optimal policy is guaranteed to be exactly optimal for sufficiently small ϵ because Π is finite. VI's running time-complexity is polynomial in $|X|, |A|, 1/(1-\gamma)$, and the size of representing the inputs R and P (Blondel & Tsitsiklis, 2000).

VI is known to be one of the two exact dynamic-programming methods along with policy iteration (PI) (Puterman, 1994). Due to its simplicity and the exactness but the dimensional non-scalability, since VI was developed by Bellman (1957), a great body of works has been done to implement it in real applications and to improve or approximate it (to have an approximately optimal policy) over the decades (see, e.g., Bertsekas & Tsitsiklis, 1996, Chang, Hu, Fu, & Marcus, 2013, Powell, 2011 and Puterman, 1994 and the references therein). In particular, there exist (classical) variants of VI, e.g., Jacobi, Gauss–Seidel, action-elimination, etc., (see, e.g., Puterman, 1994 and other variants therein) aiming at reducing computational complexity of VI and possibly improving

[☆] 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.

http://dx.doi.org/10.1016/j.automatica.2014.05.009
0005-1098/© 2014 Elsevier Ltd. All rights reserved.

the linear convergence rate by γ-contraction with no smaller contraction. These methods maintain the exactness of VI and are based on a single value-function manipulation. There are also some more recent efforts on devising an exact algorithm as a variant of VI to improve the convergence rate of VI by designing acceleration operator/estimator (Herzberg & Yechiali, 1994; Shlakhter, Lee, Khmelev, & Javer, 2010) or by using a sequence of truncated models (Arruda, Ourique, LaCombe, & Almudevar, 2013) with a single value-function manipulation but the degree of speeding up is not theoretically quantified even if some experimental results are provided to advocate these approaches.

In this communique, we present a novel *exact* algorithm called "value set iteration". (VSI) for solving MDPs as a simple *generalization* of VI and as a counterpart to policy set iteration (PSI) recently presented by Chang (2013). PSI is a generalization of PI by manipulating a set of policies at each iteration. As a counterpart to PSI, VSI generates a sequence of value functions based on manipulating *a set of value functions* at each iteration, as opposed to other existing exact variants of VI, and it converges to the optimal value function. VSI preserves convergence properties of VI while converging no slower than VI. In particular, if the value-function set used in VSI contains the value functions of $N \geq 1$ independently generated sample-policies from a given distribution and a properly defined policy switching policy (Chang et al., 2013), a probabilistic exponential convergence rate of VSI can be additionally established in terms of N but *independently of* γ, similar to PSI. This then potentially overcomes a major problem of the dependence on $1/(1-\gamma)$ in the running time-complexity of VI. Because the set used in VSI can contain the value functions of any policies generated by other existing algorithms, VSI is also a *general framework of combining multiple solution methods*.

We note that even if VSI manipulates a set of value functions as in PSI, PSI is based on extending the single-policy improvement step of PI into a multi-policy improvement step whereas VSI is based on a newly devised contraction-mapping operator in the space of value functions. The operator is defined for the first time in this work and iterative approximation by successive applications of the operator is totally different aspect from PSI. Each iteration of VSI requires $O((N+m+1)(|X|^2|A|+|X|^3))$ time-complexity if the set involved with VSI 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 per-iteration time-complexity of VI by a factor of about N and by the amount of evaluating the value functions in the set, no slower convergence than VI in terms of the number of iterations is guaranteed while achieving a probabilistic exponential convergence rate. We provide a finite-time probabilistic error-bound in obtaining the optimal value function for a given initial state distribution (cf., Theorem 4). One of the key ideas for the analysis is based on a probability bound of sample-maximum estimate of a random variable (Campi & Calafiore, 2009) obtained from the scenario design method (Calafiore, 2010; Calafiore & Campi, 2006) to effectively solve control design problems that can be cast in the form of a convex optimization problem with uncertain constraints. In this sense, as in PSI, VSI takes the spirit of randomized methods in probabilistic robust control.

2. Value set iteration

2.1. General framework

Let $\mathcal{P}(\Pi)$ be the power set of Π. Define a mapping $T : B(X) \times \mathcal{P}(\Pi) \to B(X)$ such that for $x \in X, u \in B(X)$, and nonempty $\Delta \in \mathcal{P}(\Pi)$,

$$T(u,\Delta)(x) := \max_{a\in A(x)}\left(R(x,a) + \gamma\sum_{y\in X} P^{a}_{xy}\max\left\{u(y), \max_{\pi\in\Delta} V^{\pi}(y)\right\}\right)$$

and $T(u,\Delta)(x) := L(u)(x)$ if $\Delta = \emptyset$.

The following lemma states that similar to L, V^* is a unique fixed point of T for any $\Delta \in \mathcal{P}(\Pi)$ and T is also a contraction mapping in $B(X)$ for any Δ. In the sequel, the norm $\|\cdot\|$ denotes $\max_{x\in X}|f(x)|$ for $f \in B(X)$ and for $u, v \in B(X)$, $u \leq (\geq) v$ means $u(x) \leq (\geq) v(x)$ for all $x \in X$.

Lemma 1. *With the mapping T, the following holds for any $\Delta \in \mathcal{P}(\Pi)$:*

1. *V^* uniquely satisfies $T(V^*, \Delta) = V^*$.*
2. *For any $u, v \in B(X)$, $\|T(u,\Delta) - T(v,\Delta)\| \leq \gamma\|u - v\|$.*

Proof. The proof of (1) is from the definitions of T and V^* and Banach's fixed point theorem. For the part (2), if $\Delta = \emptyset$, it is trivial. If $\Delta \neq \emptyset$, then for any $x \in X$ and $u, v \in B(X)$,

$$\begin{aligned}
T(u,\Delta)(x) - T(v,\Delta)(x) &\leq \gamma\sum_{y\in X} P^{a^*}_{xy}\left(\max\left\{u(y), \max_{\pi\in\Delta} V^{\pi}(y)\right\}\right. \\
&\quad \left. - \max\left\{v(y), \max_{\pi\in\Delta} V^{\pi}(y)\right\}\right) \\
&\text{where } a^* \in \arg\max_{a\in A(x)}\left(R(x,a)\right. \\
&\quad \left. + \gamma\sum_{y\in X} P^{a}_{xy}\max\left\{u(y), \max_{\pi\in\Delta} V^{\pi}(y)\right\}\right) \\
&\leq \gamma\max_{z\in X}\left|\max\left\{u(z), \max_{\pi\in\Delta} V^{\pi}(z)\right\} - \max\left\{v(z), \max_{\pi\in\Delta} V^{\pi}(z)\right\}\right| \\
&\leq \gamma\max_{z\in X}\left|u(z) - v(z)\right|.
\end{aligned}$$

Changing the role of u and v, we have that $T(v,\Delta)(x) - T(u,\Delta)(x) \leq \gamma\max_{z\in X}|u(z) - v(z)|$. This concludes $\|T(u,\Delta) - T(v,\Delta)\| \leq \gamma\|u - v\|$.

We now provide VSI below. VSI degenerates to VI if $\Delta_k = \emptyset$ for all $k \geq 0$. The structure of the algorithm follows that of VI. A sequence of the value functions $\{V_k\}$ is generated by successive applications of T with $V_0 \in B(X)$ where an arbitrary $\Delta_k \in \mathcal{P}(\Pi)$ is employed at k.

Value set iteration (VSI)

1. **Initialization**: Select $\epsilon > 0$. Set $k = 0$ and choose any $V_0 \in B(X)$.
2. **Loop**:
   2.1 Select $\Delta_k \in \mathcal{P}(\Pi)$ and obtain $V_{k+1} = T(V_k, \Delta_k)$.
   2.2 If $\|V_{k+1} - V_k\| \leq \epsilon \cdot \frac{1-\gamma}{2\gamma}$, exit the loop. Otherwise, $k \leftarrow k+1$.

The parts of (1) and (2) of the following theorem establish the similar bounds on the performance of VSI to VI's and (3) shows that VSI terminates in a finite number of iterations. That is, VSI preserves the main convergence properties of VI. The part (1) states that $\{V_k\}$ converges to V^* with a linear convergence rate of γ and the part (2) states that the policy π_k defined with V_{k+1} is ϵ-optimal. In addition, the part (4) establishes that $V_{k+1}(x)$ is lower bounded by $\max_{\pi\in\Delta_k} V^{\pi}(x)$ for all $x \in X$ so that $\|V^* - V_{k+1}\| \leq \max_{x\in X}|V^*(x) - \max_{\pi\in\Delta_k} V^{\pi}(x)|$. We will further investigate the usefulness of this property later (cf., Theorems 3 and 4). Finally, by the part (5), VSI converges to V^* no slower than VI in terms of the number of iterations.

Theorem 2. *For the sequence $\{V_k\}$ generated by VSI, and the policy π_k defined such that for all $x \in X$,*

$$\pi_k(x) \in \arg\max_{a\in A(x)}\left(R(x,a) + \gamma\sum_{y\in X} P^{a}_{xy}V_{k+1}(y)\right),$$