Contents lists available at ScienceDirect

Automatica

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

Technical communique

Value set iteration for two-person zero-sum Markov games[☆]

Hyeong Soo Chang[1]

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

ARTICLE INFO

Article history:
Received 7 October 2015
Received in revised form
30 June 2016
Accepted 8 September 2016
Available online 7 December 2016

Keywords:
Two-person zero-sum Markov game
Value iteration
Policy iteration
Stochastic game

ABSTRACT

We present a novel exact algorithm called "value set iteration" (VSI) for solving two-person zero-sum Markov games (MGs) as a generalization of value iteration (VI) and as a general framework of combining multiple solution methods. We introduce a novel operator in the value function space and iteratively apply the operator with any sequence of the set of policies, extending Chang's VSI for MDPs into the MG setting. We show that VSI for MGs converges to the equilibrium value function with at least linear convergence rate and establish that VSI can potentially improve the convergence speed in terms of the number of iterations by proper setting of the sequence of the set of policies.

1. Introduction

Consider a two-person zero-sum Markov game (MG) (see, e.g., Filar & Raghavan, 1991, Zachrisson, 1964) $M = (X, A, B, P, C)$, where X is a finite state set, and $A(x)$ and $B(x)$ are nonempty finite pure action sets for the minimizer and the maximizer, respectively, at $x \in X$ with $A = \bigcup_{x\in X} A(x)$ and $B = \bigcup_{x\in X} B(x)$. We denote the mixed action sets at $x \in X$ over $A(x)$ and $B(x)$ as $F(x)$ and $G(x)$, respectively, with $F = \bigcup_{x\in X} F(x)$ and $G = \bigcup_{x\in X} G(x)$.

Once $f \in F(x)$ and $g \in G(x)$ are simultaneously taken at $x \in X$ by the minimizer and the maximizer, respectively (with the knowledge of the state but without knowing each other's current action being taken), x makes a transition to a next state y by the probability P^{fg}_{xy} given as $P^{fg}_{xy} = \sum_{a\in A(x)} \sum_{b\in B(x)} f(a)g(b)\, p(y|x, a, b)$. Here $f(a)$ denotes the probability of selecting a, similarly $g(b)$, and $p(y|x, a, b)$ denotes the probability of transition from x to y by a and b. The minimizer then obtains an expected cost of $C(x, f, g)$ given by $C(x, f, g) = \sum_{y\in X} \sum_{a\in A(x)} \sum_{b\in B(x)} c(x, a, b)p(y|x, a, b)f(a)g(b)$, where $c(x, a, b) \in \mathbb{R}$ is a payoff to the minimizer (the negative of this is incurred to the maximizer). We let $C_{\max} = \max_{x,a,b} |c(x, a, b)|/(1 - \gamma)$. Note that the payoff function or matrix constructed from the function depends on given state.

We define a (stationary Markovian randomized) policy π of the minimizer as a mapping $\pi : X \to F$ with $\pi(x) \in F(x)$, $\forall x \in X$, and denote Π as the set of all possible such policies. A policy ϕ is similarly defined for the maximizer with G, and we denote Φ as the set of all possible such policies. Define *the value of* $\pi \in \Pi$ *and* $\phi \in \Phi$ with an initial state $x \in X$ as $V(\pi, \phi)(x) = E\left[\sum_{t=0}^{\infty} \gamma^t C(X_t, \pi(X_t), \phi(X_t)) \middle| X_0 = x\right]$, where X_t is a random variable denoting the state at time t by following π and ϕ and $\gamma \in (0, 1)$ is a discounting factor.

It is well-known that M has an equilibrium policy pair of $\pi^* \in \Pi$ and $\phi^* \in \Phi$ such that $\forall \pi \in \Pi$, $\forall \phi \in \Phi$,

$$V(\pi^*, \phi)(x) \le V(\pi^*, \phi^*)(x) \le V(\pi, \phi^*)(x)$$

and $V^* := V(\pi^*, \phi^*)$ is referred to as *the equilibrium value function of* M. We say that $\pi \in \Pi$ for the minimizer is ϵ-optimal with $\epsilon > 0$ if $\max_{x\in X} |V^*(x) - \sup_{\phi\in\Phi} V(\pi, \phi)(x)| \le \epsilon$. That is, if the equilibrium value of M at $x \in X$ is approximated by the game value obtained when the minimizer plays a policy π and the maximizer plays the best-responsive policy to π with an error at most by ϵ at every x, π is called ϵ-optimal. (The ϵ-optimality for the maximizer's case is similarly defined.)

Let $B(X)$ be the set of all real-valued functions defined on X. In the sequel, the norm $\|\cdot\|$ denotes $\max_{x\in X} |h(x)|$ for $h \in B(X)$ and for $u, v \in B(X)$, $u = (\le, \ge)v$ means $u(x) = (\le, \ge)v(x)$ for all $x \in X$. Define a mapping $L : B(X) \to B(X)$ such that for all $x \in X$ and $u \in B(X)$, $L(u)(x) := \inf_{f\in F(x)} \sup_{g\in G(x)} \left(C(x, f, g) + \gamma \sum_{y\in X} P^{fg}_{xy} u(y)\right)$. Then 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 $L(U_k) = U_{k+1}$,

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

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

[1] Fax: +82 2 704 8273.

http://dx.doi.org/10.1016/j.automatica.2016.10.010

$k \geq 0$ converges to V^*. This exact algorithm, developed by Shapley (1953), is called value iteration (VI). Since then, some variants of VI have been further developed while keeping the exactness, based on splitting the transition matrices with the similar technique used in the MDP case (cf., Theorem 6.3.7 in Puterman, 1994), to improve the linear convergence rate γ by $\alpha \in (0, 1)$ where $\alpha \leq \gamma$ (see, e.g., Filar & Raghavan, 1991, Kushner, 2004 Van Der Wal, 1977). However, all of these works have still *linear* convergence rates. More importantly, fixing a player's policy in MG induces a Markov decision process (MDP) (Puterman, 1994) and the convergence speed of VI (or the variants of VI) in this case *in terms of the number of iterations* is at least polynomial in $|X|$, $\max\{|A|, |B|\}$, $1/(1-\gamma)$, and the size of representing the inputs C and P (Blondel & Tsitsiklis, 2000). Recently, it has been also shown that the number of iterations can be exponential with respect to the number of actions (Feinberg & Huang, 2014). To the author's best knowledge, there has been no notable work of developing an *exact* algorithm that addresses the issues of the linear convergence rate and the crucial dependence on $O(1/(1-\gamma))$ in the convergence speed of VI in terms of the number of iterations. By simply taking γ close to 1, the number of iterations to converge becomes almost infinite.

In this communique, we present a novel exact algorithm called "value set iteration" (VSI) for solving two-person zero-sum MGs as a generalization of VI and as a general framework of combining multiple solution methods. We introduce a novel operator in the value function space and iteratively apply the operator with any sequence of the set of policies, extending VSI for MDPs (Chang, 2014) into the MG setting. We show that VSI for MGs converges to the equilibrium value function, V^*, with at least linear convergence rate and establish that VSI can potentially improve the convergence speed in terms of the number of iterations by proper setting of the sequence of the set of policies. Our focus is on the convergence property but not on solving the well-known curse of dimensionality problem of VI. Even if VSI subsumes the properties of VI, it also still suffers from the dimensionality problem.

VSI works with a sequence $\{\Delta_k\}$ of the sets of the policies (of a player). At iteration k of VSI, V^* is first "estimated" by using the value functions of the games when each policy in Δ_k is played and its corresponding best-responsive policy is played by the rivalry. The estimate function is then "tuned" by comparing it with the current estimate function of VSI, V_k, for V^*, which is in turn used for finally generating the estimate function of VSI, V_{k+1}, for V^* at $k+1$. Because any policy can be an element of Δ_k, $\{\Delta_k\}$ provides considerable freedom of designing a *convergent exact* algorithm, which potentially allows to overcome the dependence on $1/(1-\gamma)$ in the time-complexity. In other words, a proper choice of $\{\Delta_k\}$ leads to an algorithm whose convergence rate is faster than the linear convergence rate of VI. Furthermore, VSI serves as a general framework of combining multiple solution methods for MGs by the role of $\{\Delta_k\}$. For example, VSI can incorporate the policies generated by policy iteration (PI) (Rao, Chandrasekaran, & Nair, 1973; Van Der Wal, 1995) or the other variants of VI, etc.

Adding some functional into VI for VSI makes VI's time-complexity *per iteration* in terms of the number of the algebraic operations increase. This is mainly from solving $|\Delta_k|$ independent MDPs at iteration k. If these MDPs can be computed in parallel, VSI has a similar time-complexity per iteration to that of PI.

2. Value set iteration

The exposition of the algorithm below is in the perspective of the minimizer. The maximizer's case can be symmetrically given. Let $\mathcal{P}(\Pi)$ be the set of all *finite* subsets of Π. Define a mapping $T : B(X) \times \mathcal{P}(\Pi) \to B(X)$ such that for all $x \in X$, $u \in B(X)$, and nonempty $\Delta \in \mathcal{P}(\Pi)$,

$$T(u, \Delta)(x) := \inf_{f \in F(x)} \sup_{g \in G(x)} \Bigg(C(x, f, g) + \gamma \sum_{y \in X} P_{xy}^{fg} \min\Big\{ u(y), \min_{\pi \in \Delta} \sup_{\phi \in \Phi} V(\pi, \phi)(y) \Big\} \Bigg)$$

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

The lemma below 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.

Lemma 1. *The following holds for any $\Delta \in \mathcal{P}(\Pi)$ by T:*

1. *For any $u, v \in B(X)$, $\|T(u, \Delta) - T(v, \Delta)\| \leq \gamma \|u - v\|$.*
2. *V^* uniquely satisfies $T(V^*, \Delta) = V^*$.*
3. *For any $u, v \in B(X)$ such that $u \leq v$, $T(u, \Delta) \leq T(v, \Delta)$.*
4. *For any $u \in B(X)$ and $c > 0$, $T(u + c, \Delta) \leq T(u, \Delta) + c\gamma$ and $T(u - c, \Delta) \geq T(u, \Delta) - c\gamma$, where $u + c(x) := u(x) + c$ and $u - c(x) := u(x) - c$, $\forall x \in X$.*

Proof. For the part (1), if $\Delta = \emptyset$, it is trivial. If $\Delta \neq \emptyset$, then for any $u, v \in B(X)$,

$$\begin{aligned}
&\|T(u, \Delta) - T(v, \Delta)\| \\
&\leq \gamma \max_{x \in X} \sup_{f,g} \Bigg| \sum_{y \in X} P_{xy}^{fg} \min\Big\{u(y), \min_{\pi \in \Delta} \sup_{\phi \in \Phi} V(\pi, \phi)(y)\Big\} \\
&\qquad - \min\Big\{v(y), \min_{\pi \in \Delta} \sup_{\phi \in \Phi} V(\pi, \phi)(y)\Big\} \Bigg| \\
&\leq \gamma \max_{x \in X} \sup_{f,g} \sum_{y \in X} P_{xy}^{fg} \Bigg| \min\Big\{u(y), \min_{\pi \in \Delta} \sup_{\phi \in \Phi} V(\pi, \phi)(y)\Big\} \\
&\qquad - \min\Big\{v(y), \min_{\pi \in \Delta} \sup_{\phi \in \Phi} V(\pi, \phi)(y)\Big\} \Bigg| \\
&\leq \gamma \max_{x \in X} \sup_{f,g} \sum_{y \in X} P_{xy}^{fg} \max_{z \in X} \Bigg| \min\Big\{u(z), \min_{\pi \in \Delta} \sup_{\phi \in \Phi} V(\pi, \phi)(z)\Big\} \\
&\qquad - \min\Big\{v(z), \min_{\pi \in \Delta} \sup_{\phi \in \Phi} V(\pi, \phi)(z)\Big\} \Bigg| \\
&\leq \gamma \max_{x \in X} \sup_{f,g} \sum_{y \in X} P_{xy}^{fg} \max_{z \in X} |u(z) - v(z)| = \gamma \|u - v\|.
\end{aligned}$$

The proof of (2) is from the definitions of T and V^*, where we use $V^*(x) = \inf_{\pi \in \Pi} \sup_{\phi \in \Phi} V(\pi, \phi)(x)$, $\forall x \in X$, and from the fixed point theorem. The part (3), referred to as the monotonicity property, follows easily from the assumption $u \leq v$ and the definition of T. The proof of (4) comes from the fact that for any $h, i \in B(X)$ and $c > 0$, $\min\{h(x) + c, i(x)\} \leq \min\{h(x), i(x)\} + c$ and $\min\{h(x) - c, i(x)\} \geq \min\{h(x), i(x)\} - c$ for all $x \in X$.

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 $\|V_{k+1} - V_k\| \leq \epsilon \cdot \frac{(1-\gamma)^2}{2\gamma}$, exit the loop. Otherwise, $k \leftarrow k + 1$.

We need to define more operators to study the properties of VSI: Define a mapping $L_\pi : B(X) \rightarrow B(X)$ for $\pi \in \Pi$ such that for all $x \in X$ and $u \in B(X)$,

$$L_\pi(u)(x) = \sup_{g \in G(x)} \left(C(x, \pi(x), g) + \gamma \sum_{y \in X} P_{xy}^{\pi(x)g} u(y) \right)$$

and a mapping $T_{\pi,*} : B(X) \times \mathcal{P}(\Pi) \rightarrow B(X)$ for $\pi \in \Pi$ such that for all $x \in X$, $u \in B(X)$, and nonempty $\Delta \in \mathcal{P}(\Pi)$,

$$T_{\pi,*}(u, \Delta)(x) := \sup_{g \in G(x)} \Bigg(C(x, \pi(x), g) + \gamma \sum_{y \in X} P_{xy}^{\pi(x)g} \min\left\{ u(y), \min_{\pi' \in \Delta} \sup_{\phi \in \Phi} V(\pi', \phi)(y) \right\} \Bigg)$$

and $T_{\pi,*}(u, \Delta) := L_\pi(u)$ if $\Delta = \emptyset$. Note that $T_{\pi,*}$-operator also has the similar properties as stated in Lemma 1.

We remark that at iteration $k \geq 0$ PI obtains π_{k+1} for a given $W_k \in B(X)$ such that $T_{\pi_{k+1},*}(W_k, \phi) = T(W_k, \phi)$ in the policy improvement step and obtains W_{k+1} such that $W_{k+1}(x) = \sup_{\phi \in \Phi} V(\pi_{k+1}, \phi)(x)$ for all $x \in X$ in the value evaluation step, where W_0 is selected such that $T(W_0) \leq W_0$. On the other hand, in VSI, $|\Delta_k|$ independent MDPs induced by fixing each $\pi \in \Delta_k$ need to be solved at k. If these MDPs can be computed in parallel (or $\max_k |\Delta_k|$ is relatively small), the complexity of computing $\min_{\pi \in \Delta_k} \sup_{\phi \in \Phi} V(\pi, \phi)(y)$, $y \in X$, is of the same order as that of the value evaluation step of PI. Therefore, the time-complexity per iteration of VSI would be of the same order as that of PI.

The part (1) of the following theorem establishes the similar bound on the performance of VSI to VI's and the part (2) shows that VSI terminates in a finite number of iterations. The part (3) establishes that the policy π_k defined greedily with respect to V_{k+1} for the minimizer is ϵ-optimal when VSI terminates. In addition, the part (4) establishes that V_{k+1} upper bounds V^* and $V_{k+1}(x)$ is upper bounded by $\min_{\pi \in \Delta_k} \sup_{\phi \in \Phi} V(\pi, \phi)(x)$ for all $x \in X$ so that $\|V^* - V_{k+1}\| \leq \max_{x \in X} |V^*(x) - \min_{\pi \in \Delta_k} \sup_{\phi \in \Phi} V(\pi, \phi)(x)|$. We can see that the degree of estimation by V_{k+1} for V^* can be tuned by policies in Δ_k, independently of the value of γ. We will further investigate this property later (cf., Theorem 3). 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 $L_{\pi_k}(V_{k+1}) = L(V_{k+1})$, and the sequence $\{U_k\}$ generated by VI with $k \geq 0$, the following holds for any $\{\Delta_k\}$:*

1. $\|V^* - V_k\| \leq \frac{2\gamma^k C_{\max}}{1-\gamma}$ *for* $k \geq 0$.
2. *There exists $N < \infty$ such that $\|V_{k+1} - V_k\| \leq \epsilon \cdot \frac{(1-\gamma)^2}{2\gamma}$ for all $k \geq N$.*
3. *The policy π_k is ϵ-optimal when VSI terminates, i.e., $\max_{x \in X} |V^*(x) - \sup_{\phi \in \Phi} V(\pi_k, \phi)(x)| \leq \epsilon$.*
4. *If $V_0(x) = \sup_{\phi \in \Phi} V(\pi, \phi)(x)$ for all $x \in X$ for some $\pi \in \Pi$, then for all $x \in X$ and $k \geq 0$,*
$$V^*(x) \leq V_{k+1}(x) \leq \min_{\pi \in \Delta_k} \sup_{\phi \in \Phi} V(\pi, \phi)(x).$$
5. *If $V_0 = U_0$, then $\|V^* - V_k\| \leq \|V^* - U_k\|$ for $k \geq 0$.*

Proof. Throughout the proof, fix any $\{\Delta_k\}$. Because the results extend the well-known results of VI, some parts of the proof basically follow the standard techniques in the literature.

The statement (1) directly comes from the contraction property of T stated in Lemma 1. $\|V^* - V_k\| = \|T(V^*, \Delta_{k-1}) - T(V_{k-1}, \Delta_{k-1})\| \leq \gamma \|V^* - V_{k-1}\| \leq \gamma^2 \|V^* - V_{k-2}\| \leq \cdots \leq \gamma^k \|V^* - V_0\| \leq 2\gamma^k C_{\max}/(1-\gamma)$.

For the part (2), from $\|V^* - V_k\| \leq \gamma^k \|V^* - V_0\|$ and the boundedness of V^* and V_0, for any $\alpha > 0$, there exists $N < \infty$ such that for all $k \geq N$, $\|V^* - V_k\| \leq \alpha$. Then $V_{k+1} = T(V_k, \Delta_k) \leq T(V^* + \alpha, \Delta_k) \leq T(V^*, \Delta_k) + \alpha\gamma = V^* + \alpha\gamma$, where the last inequality from the part (4) of Lemma 1. Similarly $V_{k+1} \geq V^* - \alpha\gamma$. We have that $\|V_{k+1} - V^*\| \leq \alpha\gamma$ for all $k \geq N$. Therefore, by setting $\alpha = \epsilon(1-\gamma)^2/(2\gamma(1+\gamma))$, we have that $\|V_{k+1} - V_k\| \leq \|V_{k+1} - V^*\| + \|V^* - V_k\| \leq \alpha(1+\gamma) = \epsilon(1-\gamma)^2/(2\gamma)$ for all $k \geq N$.

For the part (3), let $V^{\pi_k}(x) = \sup_{\phi \in \Phi} V(\pi_k, \phi)(x)$ for all $x \in X$. We have that

$$\|V^* - V^{\pi_k}\| \leq \|V^* - V_{k+1}\| + \|V_{k+1} - V^{\pi_k}\|. \tag{1}$$

For the first term of $\|V^* - V_{k+1}\|$ in the previous inequality,

$$\begin{aligned} \|V^* - V_{k+1}\| &\leq \|V^* - T(V_{k+1}, \Delta_k)\| + \|T(V_{k+1}, \Delta_k) - V_{k+1}\| \\ &= \|T(V^*, \Delta_k) - T(V_{k+1}, \Delta_k)\| \\ &\quad + \|T(V_{k+1}, \Delta_k) - T(V_k, \Delta_k)\| \\ &\leq \gamma \|V^* - V_{k+1}\| + \gamma \|V_{k+1} - V_k\| \quad \text{by Lemma 1.} \end{aligned}$$

It follows that

$$\|V^* - V_{k+1}\| \leq \frac{\gamma}{1-\gamma} \|V_{k+1} - V_k\|. \tag{2}$$

For the second term, $\|V^{\pi_k} - V_{k+1}\| \leq \|L_{\pi_k}(V^{\pi_k}) - L(V_{k+1})\| + \|L(V_{k+1}) - V_{k+1}\|$ by using $V^{\pi_k} = L_{\pi_k}(V^{\pi_k})$ from the MDP induced by π_k. The right hand side of the inequality is then equal to $\|L_{\pi_k}(V^{\pi_k}) - L_{\pi_k}(V_{k+1})\| + \|L(V_{k+1}) - V_{k+1}\|$ by using $L(V_{k+1}) = L_{\pi_k}(V_{k+1})$. Therefore,

$$\begin{aligned} &\|V^{\pi_k} - V_{k+1}\| \\ &\quad \leq \gamma \|V^{\pi_k} - V_{k+1}\| + \|L(V_{k+1}) - V^*\| + \|V^* - V_{k+1}\| \\ &\quad = \gamma \|V^{\pi_k} - V_{k+1}\| + \|L(V_{k+1}) - L(V^*)\| + \|V^* - V_{k+1}\| \\ &\quad \leq \gamma \|V^{\pi_k} - V_{k+1}\| + \gamma \|V_{k+1} - V^*\| + \|V^* - V_{k+1}\| \\ &\quad = \gamma \|V^{\pi_k} - V_{k+1}\| + (1+\gamma) \|V^* - V_{k+1}\|. \end{aligned}$$

Therefore, we have that

$$\|V^{\pi_k} - V_{k+1}\| \leq \frac{1+\gamma}{1-\gamma} \cdot \|V^* - V_{k+1}\| \leq \frac{\gamma(1+\gamma)}{(1-\gamma)^2} \cdot \|V_{k+1} - V_k\|. \tag{3}$$

By combining the bounds of (2) and (3) into (1), it finally follows that $\|V^* - V^{\pi_k}\| \leq \frac{2\gamma}{(1-\gamma)^2} \|V_{k+1} - V_k\| \leq \epsilon$.

We now prove the part (4). If $V_0(x) = \sup_{\phi \in \Phi} V(\pi, \phi)(x)$ for some $\pi \in \Pi$, it is trivial that $V^* \leq V_0$. Assume that for $k = n$, $V^* \leq V_n$, Then $V^* = T(V^*, \Delta_n) \leq T(V_n, \Delta_n) = V_{n+1}$. Therefore, $V^* \leq V_k$ for all $k \geq 0$. Now, for any $\pi' \in \Delta_k$ and $x \in X$,

$$\begin{aligned} &V_{k+1}(x) \leq T_{\pi',*}(V_k, \Delta_k)(x) \\ &\quad \leq \sup_{g \in G(x)} \left(C(x, \pi'(x), g) + \gamma \sum_{y \in X} P_{xy}^{\pi'(x)g} \sup_{\phi' \in \Phi} V(\pi', \phi')(y) \right) \\ &\quad = \sup_{\phi \in \Phi} V(\pi', \phi)(x), \end{aligned}$$

where again the last equality follows from Bellman's optimality equation for the MDP induced from the MG by fixing π'. It follows that for all $x \in X$, $V_{k+1}(x) \leq \min_{\pi \in \Delta_k} \sup_{\pi \in \Phi} V(\pi, \phi)(x)$. Finally, the statement (5) can be proven by showing that $V_k \leq U_k$ for all k by the induction on k.

Because VSI converges with any sequence of $\{\Delta_k\}$, it provides a considerable freedom of designing convergent variants of VI. Recall that $\|V^* - V_{k+1}\| \leq \max_{x \in X} |V^*(x) - \min_{\pi \in \Delta_k} \sup_{\phi \in \Phi} V(\pi, \phi)(x)|$. In the next section, we show that this bound can be made tighter by properly designing $\{\Delta_k\}$.

3. Value set iteration with policy switching

We define the *policy switching* policy $\pi_{\text{ps}}(\Delta)$ for a given nonempty $\Delta \in \mathcal{P}(\Pi)$ such that for all $x \in X$,

$$\pi_{\text{ps}}(\Delta)(x) \in \left\{ \pi(x) \middle| \pi \in \operatorname*{argmin}_{\pi' \in \Delta} \sup_{\phi \in \Phi} V(\pi', \phi)(x) \right\}.$$

Then the sequence $\{\Delta_k\}$ is selected such that for $k \geq 0$, $\Delta_k = \{\pi_{\text{ps}}(\Delta_{k-1})\} \cup \Psi_k$, where $\Psi_k \in \mathcal{P}(\Pi)$ and $\{\pi_{\text{ps}}(\Delta_{-1})\} := \emptyset$.

VSI with Policy Switching (VSI-PS)

1. **Initialization:** Select $\epsilon > 0$. Set $k = 0$ and $\{\pi_{\text{ps}}(\Delta_{-1})\} := \emptyset$. Select any $V_0 \in B(X)$.
2. **Loop:**
   2.1 $\Delta_k \leftarrow \{\pi_{\text{ps}}(\Delta_{k-1})\} \cup \Psi_k$, where $\Psi_k \in \mathcal{P}(\Pi)$.
   2.2 Obtain $V_{k+1} = T(V_k, \Delta_k)$.
   2.3 If $\|V_{k+1} - V_k\| \leq \epsilon \cdot \frac{(1-\gamma)^2}{2\gamma}$, exit the loop. Otherwise, $k \leftarrow k + 1$.

Theorem 3. *For the sequence $\{V_k\}$ generated by VSI-PS, suppose that $V_0(x) = \sup_{\phi \in \Phi} V(\pi, \phi)(x)$ for all $x \in X$ for some $\pi \in \Pi$. Then for all $k \geq 0$ and $x \in X$,*

$$V^*(x) \leq V_{k+1}(x) \leq \min_{\pi \in \bigcup_{k'=0}^{k} \Delta_{k'}} \sup_{\phi \in \Phi} V(\pi, \phi)(x).$$

Proof. It is obvious that $V^* \leq V_0$. Assume that $V^* \leq V_k$. Then $V_{k+1} = T(V_k, \Delta_k) \geq T(V^*, \Delta_k) = V^*$. Therefore for all k, $V_k \geq V^*$.

From Theorem 2(4), $\min_{\pi \in \Delta_k} \sup_{\phi \in \Phi} V(\pi, \phi)(x) \geq V_{k+1}(x)$ for all $x \in X$. Observe that for all $x \in X$, $\sup_{\phi \in \Phi} V(\pi_{\text{ps}}(\Delta_k), \phi)(x) \leq \min_{\pi \in \Delta_k} \sup_{\phi \in \Phi} V(\pi, \phi)(x)$ from the fact established in the MDP setting (Chang, Hu, Fu, & Marcus, 2013, Chp. 3): Given $\Delta \in \mathcal{P}(\Pi)$, we have that for any fixed policy $\phi \in \Phi$, $V(\xi, \phi)(x) \leq \min_{\pi \in \Delta} V(\pi, \phi)(x)$ for all $x \in X$ if ξ is defined as $\xi(x) \in \left\{ \pi(x) \middle| \pi \in \operatorname{argmin}_{\pi' \in \Delta} V(\pi', \phi)(x) \right\}$, $\forall x \in X$.

Because $\pi_{\text{ps}}(\Delta_{k-1}) \in \Delta_k$, we have that $\min_{\pi \in \Delta_k \cup \Delta_{k-1}} \sup_{\phi \in \Phi} V(\pi, \phi)(x) \geq V_{k+1}(x)$ for all $x \in X$. Continuing this way, we have that for all $x \in X$, $\min_{\pi \in \bigcup_{k'=0}^{k} \Delta_{k'}} \sup_{\phi \in \Phi} V(\pi, \phi)(x) \geq V_{k+1}(x)$.

We remark that all of the results in Theorem 2 apply to VSI-PS. In particular, the following result is immediate and summarizes the convergence speed of VSI-PS:

Theorem 4. *Given any sequence $\{\Delta_k\}$, for $k \geq 0$ and $\{V_k\}$ generated by VSI-PS,*

$$\|V^* - V_{k+1}\| \leq \min\left\{ \frac{2\gamma^{k+1} C_{\max}}{1-\gamma}, \|V^* - \Phi_k\| \right\},$$

where $\Phi_k(x) := \min_{\pi \in \bigcup_{k'=0}^{k} \Delta_{k'}} \sup_{\phi \in \Phi} V(\pi, \phi)(x), x \in X$.

For the case of VSI, $\min_{\pi \in \bigcup_{k'=0}^{k} \Delta_{k'}}$ is simply replaced with $\min_{\pi \in \Delta_k}$. We can see that the convergence speed of VSI-PS (or VSI) can be improved by a proper choice of $\{\Delta_k\}$.

4. Concluding remark

A future work is to study characterizing the choice of $\{\Delta_k\}$ and its corresponding convergence rate. Theoretical and empirical studies about the convergence rate with various choices of $\{\Delta_k\}$ are important. A possible choice of $\{\Delta_k\}$ in VSI ($\{\Psi_k\}$ in VSI-PS) is to make it contain the policies generated by PI, possibly with other policies by (heuristic) approximation approaches. There exists a sufficient condition (Puterman, 1994, Theorem 6.4.8) for a quadratic convergence rate of PI for MDPs. We can consider studying extension of the result having a sufficient condition for a quadratic convergence rate of VSI. If a quadratic convergence rate of PI in the MG setting is possible and if we put those policies generated by PI into $\{\Delta_k\}$ in VSI, then VSI's convergence rate would be also quadratic. Another example of setting up $\{\Delta_k\}$ is to sample policies uniformly over the policy space. In the MDP case (Chang, 2014), it was shown that the corresponding version has a "probabilistic exponential" convergence rate. Randomly generated sample sequence of $\{\Delta_k\}$ and its relationship with the performance of VSI is a good topic to inspect.

Note that the main focus of this note is on improving the convergence speed of VI in terms of the number of iterations but not on the per-iteration time-complexity. VSI still has the curse of dimensionality problem if $|X|$ and/or $|A|$ and/or $|B|$ is large. In general, employing other methods to choose policies in $\{\Delta_k\}$ would increase the per-iteration complexity. To apply VSI, it is assumed that the sizes of the state and the action spaces are not large.

The exposition of VSI was done in the perspective of the minimizer. Consider the maximizer's case as follows. We obtain the sequence $\{\tilde{V}_k\}$ such that $\tilde{V}_{k+1} = \tilde{T}(\tilde{V}_{k+1}, \tilde{\Delta}_k)$ where $\tilde{\Delta}_k \in \mathcal{P}(\Phi)$ and $\tilde{T}$ is symmetrically defined: for $u \in B(X)$, $x \in X$, and $\Delta \in \mathcal{P}(\Phi)$, $\tilde{T}(u, \Delta)(x) := \sup_{g \in G(x)} \inf_{f \in F(x)} (C(x, f, g) + \gamma \sum_{y \in X} P_{xy}^{fg} \max\{u(y), \max_{\phi \in \Delta} \inf_{\pi \in \Pi} V(\pi, \phi)(y)\})$ and we have that $\tilde{T}(u, \emptyset) := \sup_{g \in G(x)} \inf_{f \in F(x)} (C(x, f, g) + \gamma \sum_{y \in X} P_{xy}^{fg} u(y))$ with the termination condition of $\|\tilde{V}_{k+1} - \tilde{V}_k\| \leq \epsilon(1-\gamma)^2/(2\gamma)$. Then for the policy ϕ_k defined greedily with respect to $\tilde{V}_{k+1}$ for the maximizer is ϵ-optimal when VSI terminates. From this we can obtain $\|V^* - V(\pi_k, \phi_k)\| \leq \epsilon$. This is because for all $x \in X$, $-\epsilon \leq V^*(x) - \sup_{\phi \in \Phi} V(\pi_k, \phi)(x) \leq V^*(x) - V(\pi_k, \phi_k)(x)$ and $V^*(x) - V(\pi_k, \phi_k)(x) \leq V^*(x) - \inf_{\pi \in \Pi} V(\pi, \phi_k)(x) \leq \epsilon$.

References

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

Chang, H. S. (2014). Value set iteration for Markov decision processes. *Automatica*, *50*, 1940–1943.

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

Feinberg, E. A., & Huang, J. (2014). The value iteration algorithm is not strongly polynomial for discounted dynamic programming. *Operations Research Letters*, *42*, 130–131.

Filar, J. A., & Raghavan, T. E. S. (1991). Algorithms for stochastic games: A survey. *ZOR - Methods and Models of Operations Research*, *35*, 437–472.

Kushner, H. J. (2004). The Gauss–Seidel numerical procedure for Markov stochastic games. *IEEE Transactions on Automatic Control*, *49*, 1779–1782.

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

Rao, S. S., Chandrasekaran, R., & Nair, K. P. K. (1973). Algorithms for discounted games. *Journal of Optimization Theory and Applications*, *11*, 627–637.

Shapley, L. S. (1953). Stochastic games. *Proceedings of the National Academy of Sciences*, *39*, 1095–1100.

Van Der Wal, J. (1977). Discounted Markov games; successive approximation and stopping times. *International Journal of Game Theory*, *6*, 11–12.

Van Der Wal, J. (1995). Discounted Markov games: generalized policy iteration method. *Journal of Optimization Theory and Applications*, *25*, 125–138.

Zachrisson, L. E. (1964). Markov games. *Annals of Mathematics Studies*, *52*, 211–253.