Approximate value iteration with randomized policies

D.P. de Farias
Stanford University
pucci@stanford.edu

B. Van Roy
Stanford University
bvr@stanford.edu

Abstract

The curse of dimensionality in dynamic programming prevents in most problems of practical interest the exact computation of the value function. In this paper, we study the fixed points of approximate value iteration, a simple algorithm that combats the curse of dimensionality by generating approximate iterates of the classical value iteration algorithm in the span of a set of prespecified basis functions. We show that, in general, the modified dynamic programming operator need not possess a fixed point, and therefore, approximate value iteration should not be expected to converge. However, by using a class of randomized policies, approximate value iteration *is* guaranteed to possess at least one fixed point. We finally discuss the link between approximate value iteration and temporal–difference learning (TD), and show that the existence of fixed points for approximate value iteration implies existence of stationary points for the ordinary differential equation approximated by a version of TD that incorporates "exploration."

1 Introduction

Value iteration offers a simple approach to computing the value function for finite–state discounted dynamic programs. The algorithm can be described in terms of the "dynamic programming operator" T. In particular, it generates a sequence of functions according to $J_{k+1} = TJ_k$, each mapping states to real numbers. This sequence converges to the optimal value function J^*, the unique fixed point of T that can be used to generate an optimal policy. However, due to the "curse of dimensionality," for problems of practical scale, the computational burden associated with storing and manipulating functions over the state space is prohibitive, and approximations are called for. One simple method – dating all the way back to [1] – is approximate value iteration, which aims at approximating each iterate J_k by a linear combination of prespecified basis functions $\phi_1, \ldots, \phi_K$. In rough terms, iterates $\tilde{J}_k$ are given by $\tilde{J}_{k+1} = \Pi T \tilde{J}_k$, where Π is a projection operator that produces a function in the span of $\phi_1, \ldots, \phi_K$ and close to $T\tilde{J}_k$. The hope is that $\tilde{J}_k$ converges to a good approximation of J^*.

A fundamental question is whether the composition ΠT possesses a fixed point $\tilde{J}$ that may serve as a limit to the sequence $\tilde{J}_k$. It turns out – as will be illustrated by examples in Section 3 – that ΠT does not always have a fixed point. In subsequent sections, we propose and analyze a variant of approximate value iteration that *is* guaranteed to have a fixed point.

The variant of approximate value iteration developed in this paper was motivated by studies of temporal–difference learning (TD), a class of algorithms that can be viewed as simulation–based versions of approximate value iteration [3, 4, 6, 7, 8, 9]. As we will discuss, existence of fixed points for the proposed variant of approximate value iteration implies existence of stationary points for a version of TD that incorporates "exploration." Our analysis of approximate value iteration therefore also resolves an open question concerning TD.

2 Exact and Approximate Value Iteration

We consider a controlled Markov chain with a finite set of states $\mathcal{S}$ and finite sets of actions $\mathcal{A}_x, x \in \mathcal{S}$. Each state–action pair $x \in \mathcal{S}$ and $a \in \mathcal{A}_x$ is associated with a reward $g_a(x)$ and transition probabilities $P_a(x, \cdot)$. Time–relative preferences are defined by a discount factor $\alpha \in (0, 1)$. We denote by P_a a matrix whose (x, y)th component is $P_a(x, y)$, and we let $P_a(x)$ be a row vector equal to the xth row of P_a.

A stochastic stationary policy is a mapping $\mu : \{(x, a) | \ x \in \mathcal{S}, \ a \in \mathcal{A}_x\} \mapsto [0, 1]$, with $\sum_{a \in \mathcal{A}_x} \mu(x, a) = 1$ for all x. The policy defines probabilities with which actions are selected at each state. When controlled by a policy μ, the system evolves as a Markov chain with transition probabilities $P_\mu(x, y) = \sum_{a \in \mathcal{A}_x} \mu(x, a) P_a(x, y)$ and rewards $g_\mu(x) = \sum_{a \in \mathcal{A}_x} \mu(x, a) g_a(x)$. A policy μ is deterministic if for each $x \in S$, $\mu(x, a) = 1$ for some $a \in \mathcal{A}_x$. A policy μ is optimal if it attains the supremum of $\mathrm{E}\left[\sum_{k=0}^{\infty} \alpha^k g_\mu(x_k) \middle| x_0 = x\right]$ simultaneously for all initial states $x \in \mathcal{S}$.

We assume that P_μ is irreducible and aperiodic for every μ. Hence, each P_μ possesses a unique invariant distribution π_μ with $\pi_\mu(x) > 0$ for all x.

The value function J^* uniquely solves Bellman's equation $J = \max_\mu \{g_\mu + \alpha P_\mu J\}$. It is also the unique fixed point of the dynamic programming operator $TJ = \max_\mu \{g_\mu + \alpha P_\mu J\}$,. For each policy μ, we define an additional operator $T_\mu J = g_\mu + \alpha P_\mu J$. For any J, there is a deterministic policy μ such that $TJ = T_\mu J$. We call such a μ a *greedy policy* with respect to J. A policy μ^* is optimal if and only if it is greedy with respect to the optimal value function J^*.

Value iteration computes improving approximations to the value function by generating $J_{k+1} = TJ_k$. It is well–known that, for any J_0, the sequence J_k converges to J^*. Unfortunately, due to the curse of dimensionality, application of value iteration becomes infeasible in the face of problems of practical scale.

Approximate value iteration aims at alleviating the prohibitive computational burden associated with value iteration by dealing with parameterized approximations rather than functions over the state space. Given a preselected collection $\phi_1, \ldots, \phi_K$ of basis functions, the algorithm generates approximations $\tilde{J}_k = \Phi r_k$ to each iterate J_k, where $r_k \in \Re^K$ and $\Phi = [\phi_1 \cdots \phi_K]$. We denote $\phi(x) = (\phi_1(x), \ldots, \phi_K(x))'$ so that $\tilde{J}_k(x) = \phi'(x) r_k$. We assume, without loss of generality, that the basis functions are linearly independent.

The simplest form of approximate value iteration involves a projection matrix Π that projects onto the span of $\phi_1, \ldots, \phi_K$ with respect to the standard Euclidean norm, i.e., $\Pi J = \underset{\Phi r}{\operatorname{argmin}} \|J - \Phi r\|_2$. The algorithm then generates iterates according to $\tilde{J}_{k+1} = \Pi T \tilde{J}_k$. Hence, the operator T is effectively approximated by ΠT.

Alternatively, one can use projections with respect to a weighted Euclidean norm. In particular, given a vector $\pi \in \Re^{|\mathcal{S}|}$ with positive components, we define $\|J\|_\pi = \left(\sum_{x \in \mathcal{S}} \pi(x) J^2(x)\right)^{1/2}$, and the associated projection is given by $\Pi J = \underset{\Phi r}{\operatorname{argmin}} \|J - \Phi r\|_\pi$. As before, approximate value iteration generates iterates according to $\tilde{J}_{k+1} = \Pi T \tilde{J}_k$. One motivation for employing a weighted Euclidean norm is that it enables "emphasis" of "important" or "frequently visited" states in trading–off error among states.

Finally, one might not employ a single projection matrix, but rather choose a projection matrix based on the current iterate. For example, given a mapping that defines for each function J a projection matrix Π_J, iterates could be generated by $\tilde{J}_{k+1} = \Pi_{\tilde{J}_k} T \tilde{J}_k$.

3 On the Nonexistence of Fixed Points

It turns out that operators associated with simple versions of approximate value iteration often lack fixed points. In such cases, iterates $\tilde{J}_k$ will not converge to a fixed point as desired. We provide in this section two examples that illustrate difficulties that arise and motivate the variant of approximate value iteration proposed in the next section. The first example uses projections with respect to the standard Euclidean norm. In this example, the composition ΠT does not possess a fixed point and approximate value iteration leads to an unbounded sequence of iterates.

Example 3.1 *Consider an autonomous Markov chain with two states. In state 1, the reward is 1 and there is a probability 0.2 of remaining in that state and 0.8 of going to state 2. In state 2, the reward is 2 and there is a probability 0.2 of going to state 1 and 0.8 of staying in state 2. Let g be the vector of rewards and P be the transition probability matrix. The discount factor is $\alpha = \frac{5}{5.4}$ and $\Phi = [1\ 2]'$. Then if Π is the projection with respect to the Euclidean norm, we have $\Phi r = \Pi T \Phi r \Leftrightarrow r = r + 1$. Hence, ΠT does not have a fixed point. Furthermore, the sequence of weights r_k generated by approximate value iteration evolves according to $r_{k+1} = 1 + r_k$, and therefore the sequence is unbounded.*

The following lemma from [9, 10] motivates the use of a certain weighted Euclidean norm in order to circumvent the difficulty that arises in the previous example.

Lemma 3.1 *Let P be the transition matrix for an irreducible aperiodic Markov chain, and let π be the invariant distribution. Then, $\|P\|_\pi \leq 1$.*

It follows from this lemma is that, for autonomous systems (i.e., $|\mathcal{A}_x| = 1$ for all x), $\|TJ - T\overline{J}\|_\pi \leq \alpha \|J - \overline{J}\|_\pi$, for any J and $\overline{J}$, where π is the invariant distribution associated with the single policy. Also, since a projection matrix Π that projects with respect to $\|\cdot\|_\pi$ is nonexpansive with respect to $\|\cdot\|_\pi$, ΠT is a contraction. It follows that ΠT has a unique fixed point to which value iteration converges.

The above discussion identifies a version of value iteration that converges to a fixed point for autonomous systems. Can the essential idea be generalized to problems with multiple policies? In such cases, there are usually multiple invariant distributions to take into account, corresponding to different policies. One approach could be to project with respect to a Euclidean norm weighted by the invariant distribution associated with a greedy policy of the current iterate $\tilde{J}_k$. Let μ_J be a greedy policy with respect to J, and for any μ, let Π_μ be a projection onto the span of $\phi_1, \ldots, \phi_K$ with respect to $\|\cdot\|_{\pi_\mu}$. A version of approximate value iteration is $\tilde{J}_{k+1} = \Pi_{\mu_J} T J$.

Now let us define $HJ = \Pi_{\mu_J} TJ$ and $H_\mu J = \Pi_\mu T_\mu J$. To make this definition unambiguous, we need to es-

tablish what policy μ_J must be used if there is more than one greedy policy – in this case, we let μ_J be the randomized policy that takes each greedy action with the same probability. Does H possess a fixed point? We have already seen that in the case of a single policy the answer is affirmative. The following example, adapted from [3] shows that this might not be the case fixed points when there are multiple policies.

Example 3.2 *Consider a controlled Markov chain with three states and two deterministic policies 1 and 2, with rewards and transition matrices given by*

$$g_1 = g_2 = \left[\begin{array}{ccc} 0 & -1 & 1 \end{array} \right]',$$

$$P_1 = \left[\begin{array}{ccc} 0.2 & 0 & 0.8 \\ 0.4 & 0.6 & 0 \\ 0 & 1 & 0 \end{array} \right], \; P_2 = \left[\begin{array}{ccc} 0.2 & 0 & 0.8 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{array} \right].$$

Note that there are two possible actions at state 2, while no choices are offered at states 1 and 3. Let $\alpha = 0.99$ and $\Phi = [0\ 1\ 2]'$. For any function J, there are three possibilities for μ_J: μ_1 (with $P_{\mu_1} = P_1$), μ_2 (with $P_{\mu_2} = P_2$), and μ_3 (with $P_{\mu_3} = (P_1 + P_2)/2$). Now, a function J is a fixed point of H if and only if $\Pi_{\mu_J} T J = J$, or equivalently, $\Pi_{\mu_J} T_{\mu_J} J = J$. As shown before in the autonomous case, each composition $\Pi_{\mu_i} T_{\mu_i}$ $(i = 1, 2, 3)$ has a unique fixed point. Let us denote these fixed points by $J_1^* = \Phi r_1^*$, $J_2^* = \Phi r_2^*$, and $J_3^* = \Phi r_3^*$, respectively. It turns out that $r_1^* = -0.1647$, $r_2^* = 0.3311$ and $r_3^* = 0.1889$, and that $\mu_{J_1} = \mu_2$, $\mu_{J_2} = \mu_1$, and $\mu_{J_3} = \mu_1$. It follows that neither J_1, J_2, nor J_3, are fixed points of H, and therefore H has no fixed points.

This example raises the suspicion that nonexistence of fixed points may be a consequence of discontinuities of H at points where there is more than one greedy policy. As we will show, using randomized policies leads to a continuous variant of H for which fixed points are guaranteed to exist.

4 Incorporating Exploration

We now introduce a modified dynamic programming operator. This definition makes use of δ–greedy policies, which effectively incorporate exploration into a greedy policy. Formally, for any $\delta > 0$, we define a δ–greedy policy μ_J^δ with respect to J by

$$\mu_J^\delta(x, a) = \frac{\exp\left[(g_a(x) + \alpha P_a(x) J)/\delta\right]}{\sum_{\bar{a} \in \mathcal{A}_x} \exp\left[(g_{\bar{a}}(x) + \alpha P_{\bar{a}}(x) J)/\delta\right]},$$

for all $x \in \mathcal{S}$ and $a \in \mathcal{A}_x$. Our modified dynamic programming operator T_δ, which we will refer to as the δ–greedy dynamic programming operator, is then defined by $T_\delta J = T_{\mu_J^\delta} J$. Note that T_δ is continuous.

Let us now establish some basic properties of δ–greedy policies and dynamic programming operators. We first show that δ–greedy policies become greedy as $\delta \downarrow 0$.

Lemma 4.1 *Take $h \in R^m$ and let*

$$\eta^\delta(h, i) = \frac{\exp\left[h(i)/\delta\right]}{\sum_{j=1}^m \exp\left[h(j)/\delta\right]},$$

for $i = 1, \ldots, m$. Then,

$$\sup_h \left\{ \max_i h(i) - \sum_{i=1}^m \eta^\delta(h, i) h(i) \right\} \leq \frac{\delta(m-1)}{e}.$$

Proof: Without loss of generality, suppose that $h(m) = \max_i h(i)$. Then

$$\begin{aligned} &\sup_h \left\{ \max_i h(i) - \sum_{i=1}^m \eta^\delta(h, i) h(i) \right\} \\ &= \sup_h \sum_{i=1}^{m-1} \frac{(h(m) - h(i)) \exp\left[(h(i) - h(m))/\delta\right]}{1 + \sum_{j=1}^{m-1} \exp\left[(h(j) - h(m))/\delta\right]} \\ &\leq \delta \sup_h \sum_{i=1}^{m-1} \frac{h(m) - h(i)}{\delta} \exp\left[(h(i) - h(m))/\delta\right] \\ &\leq \delta(m-1) \sup_{x \geq 0} x \exp(-x) \\ &\leq \delta(m-1)/e \end{aligned}$$

■

It follows from the previous lemma that

Lemma 4.2 $\lim_{\delta \downarrow 0} \sup_{J,x} |(T_\delta J)(x) - (TJ)(x)| = 0.$

Our next lemma establishes existence of a fixed point.

Lemma 4.3 *For any $\delta > 0$, T_δ has a fixed point.*

Proof: Let $G = \sup_{x \in \mathcal{S}, a \in \mathcal{A}_x} |g_a(x)|$. Since there is only a finite number of states and actions, G is finite. Now consider the compact convex set $\{J : \|J\|_\infty \leq G/(1-\alpha)\}$. This set is closed under T_δ, and since T_δ is continuous, Brouwer's fixed point theorem guarantees existence of a fixed point. ■

Let us introduce the notion of a *quasi-contraction.*

Definition 4.1 .Quasi-contraction. *An operator F is a quasi-contraction with respect to a norm $\|\cdot\|$ if there exists a nonempty set X^* of fixed points, a compact set $\mathcal{C} \supseteq X^*$, and a scalar $\beta \in [0, 1)$ such that for any $x \notin \mathcal{C}$, there exists a fixed point $x^* \in X^*$ such that $\|Fx - x^*\| \leq \beta \|x - x^*\|$.*

Given this definition, we have the following lemma.

Lemma 4.4 *For any $\delta > 0$, T_δ is a quasi-contraction.*

Proof: For any J_1 and J_2,

$$\begin{aligned}
&\|T_\delta J_1 - T_\delta J_2\|_\infty \\
&\quad \leq \|TJ_1 - TJ_2\|_\infty + \|TJ_1 - T_\delta J_1\|_\infty + \\
&\qquad + \|TJ_2 - T_\delta J_2\|_\infty \\
&\quad \leq \alpha \|J_1 - J_2\|_\infty + O(\delta),
\end{aligned}$$

and since we know from Lemma 4.3 that T_δ has a fixed point, T_δ is indeed a quasi-contraction. ∎

The following lemma establishes that, for small δ, fixed points of T_δ approximate those of T.

Lemma 4.5 *Let J^* be the unique fixed point of T, and for any $\delta > 0$, let $\mathcal{J}^\delta$ be the set of fixed points of T_δ. Then $\lim_{\delta \downarrow 0} \sup_{J \in \mathcal{J}^\delta, x \in \mathcal{S}} |J(x) - J^*(x)| = 0$.*

Proof: For any $J \in \mathcal{J}^\delta$,

$$\begin{aligned}
\|J - J^*\|_\infty &= \|T_\delta J - J^*\|_\infty \\
&\leq \|T_\delta J - TJ\|_\infty + \|TJ - J^*\|_\infty \\
&\leq \alpha \|J - J^*\|_\infty + O(\delta),
\end{aligned}$$

and $\|J - J^*\|_\infty = O(\delta)$. The $O(\delta)$ term in the inequality is uniformly bounded over J by Lemma 4.2. ∎

Lemma 4.4 bears some important implications on T_δ. First, note that all fixed points of T_δ lie within a ball of radius $O(\delta)$. Furthermore, for J outside this circle, T_δ behaves somewhat like a contraction. Hence, for a variant of value iteration taking the form $J_{k+1} = T_\delta J_k$, after a finite number n of iterations, iterates J_k for $k \geq n$ will all lie in this ball. Furthermore, applying Lemma 4.5, we can deduce that there is some $\bar{\delta} > 0$ such that for all $\delta \leq \bar{\delta}$, greedy policies associated with functions in the ball under consideration are optimal.

5 Existence of Fixed Points

Based on the operator T_δ, we define a new version of approximate value iteration:

$$\tilde{J}_{k+1} = \Pi_{\mu^\delta_{\tilde{J}_k}} T_\delta \tilde{J}_k.$$

Alternatively, defining $H_\delta J = \Pi_{\mu^\delta_J} T_\delta J$, we have $\tilde{J}_{k+1} = H_\delta \tilde{J}_k$. The following theorem establishes that, unlike H, H_δ always possesses a fixed point.

Theorem 5.1 *For any $\delta > 0$, H_δ has a fixed point.*

To aide in the proof of this theorem we will first establish a few lemmas. Henceforth we use the shorthand notation Π^δ_r to refer to $\Pi_{\mu^\delta_{\Phi r}}$ and μ^δ_r to refer to $\mu^\delta_{\Phi r}$.

We ommit the proofs of the next two lemmas, that just establish continuity of certain functions that are important to our analysis.

Lemma 5.1 *The invariant distribution π_μ is a continuous function of μ.*

As discussed earlier, for each policy μ, there exists a unique vector r_μ such that $\Phi r_\mu = H_\mu \Phi r_\mu$ (this follows from Lemma 3.1). The next lemma establishes that the solution to this equation is continuous in μ.

Lemma 5.2 *The unique solution r_μ to $\Phi r_\mu = H_\mu \Phi r_\mu$ is a continuous function of μ.*

We now proceed to the main analysis. For any policy μ, let us define

$$s_\mu(r) = \Phi' D_\mu (T_\mu \Phi r - \Phi r), \; s_\delta(r) = \Phi' D_{\mu^\delta_r} (T_\delta \Phi r - \Phi r)$$

where $D_\mu = \text{diag}(\pi_\mu)$ for any policy μ. We also define functions $F^\gamma_\mu : \Re^K \mapsto \Re^K$ and $F^\gamma_\delta : \Re^K \mapsto \Re^K$ by

$$F^\gamma_\mu(r) = r + \gamma s_\mu(r) \quad \text{and} \quad F^\gamma_\delta(r) = r + \gamma s_\delta(r).$$

It is easy to show that fixed points of F^γ_μ and F^γ_δ coincide with those of H_μ and H_δ, as stated in the following lemma.

Lemma 5.3 *For any $\delta > 0$ and $\gamma > 0$, a vector r is a fixed point of F^γ_μ ($F^{\delta,\gamma}$) if and only if Φr is a fixed point of H_μ (H_δ).*

The next lemma establishes that, for sufficiently small γ, F^γ_μ is a pseudo–contraction.

Lemma 5.4 *There exists a constant $\gamma^* > 0$ such that for all μ and any $\gamma \in (0, \gamma^*)$, there exists a scalar $\beta_\gamma \in (0,1)$ such that*

$$\|F^\gamma_\mu(r) - r_\mu\|_2 \leq \beta_\gamma \|r - r_\mu\|_2.$$

Proof: First, note that for all μ,

$$\|H_\mu \Phi r - \Phi r_\mu\|_\mu \leq \alpha \|\Phi r - \Phi r_\mu\|_\mu,$$

and

$$\begin{aligned}
&\langle \Phi r - \Phi r_\mu, H_\mu \Phi r - \Phi r \rangle_\mu \\
&\quad = \langle \Phi r - \Phi r_\mu, (H_\mu \Phi r - \Phi r_\mu) + (\Phi r_\mu - \Phi r) \rangle_\mu \\
&\quad \leq \|\Phi r - \Phi r_\mu\|_\mu \|H_\mu \Phi r - \Phi r_\mu\|_\mu - \|\Phi r - \Phi r_\mu\|^2_\mu \\
&\quad \leq (\alpha - 1) \|\Phi r - \Phi r_\mu\|^2_\mu \\
&\quad \leq (\alpha - 1)(r - r_\mu)' (\Phi' D_\mu \Phi)(r - r_\mu).
\end{aligned}$$

Since D_μ is positive definite for all μ and the set of all randomized policies is compact, it follows that there

exists a constant $C_1 > 0$ independent of μ such that $(r - r_\mu)' s_\mu(r) \leq -C_1 \|r - r_\mu\|_2^2$. Also,

$$\begin{aligned}
\|s_\mu(r)\|^2 &= \sum_{i=1}^{K} \left(\phi_i' D_\mu (T_\mu \Phi r - \Phi r)\right)^2 \\
&= \sum_{i=1}^{K} \left(\phi_i' D_\mu (\Pi_\mu T_\mu \Phi r - \Phi r)\right)^2 \\
&\leq \sum_{i=1}^{K} \|\phi_i\|_\mu^2 \|\Pi_\mu T_\mu \Phi r - \Phi r\|_\mu^2 \\
&\leq \sum_{i=1}^{K} \|\phi_i\|_\mu^2 \left(\|\Pi_\mu T_\mu \Phi r - \Phi r_\mu\|_\mu + \|\Phi r_\mu - \Phi r\|_\mu\right)^2 \\
&\leq \sum_{i=1}^{K} \|\phi_i\|_\mu^2 \left(\alpha \|\Phi r - \Phi r_\mu\|_\mu + \|\Phi r_\mu - \Phi r\|_\mu\right)^2 \\
&= (1+\alpha)^2 \sum_{i=1}^{K} \|\phi_i\|_\mu^2 \|\Phi r_\mu - \Phi r\|_\mu^2,
\end{aligned}$$

and it follows that there exists a constant $C_2 > 0$ independent of μ such that $\|s_\mu(r)\|_2^2 \leq C_2 \|r - r_\mu\|_2^2$. Now

$$\begin{aligned}
&\|F_\mu^\gamma(r) - r_\mu\|_2^2 \\
&\quad = \|r + \gamma s_\mu(r) - r_\mu\|_2^2 \\
&\quad = \|r - r_\mu\|^2 + 2\gamma (r - r_\mu)' s_\mu(r) + \gamma^2 \|s_\mu(r)\|^2 \\
&\quad \leq (1 - 2\gamma C_1 + \gamma^2 C_2) \|r - r_\mu\|_2^2.
\end{aligned}$$

The result then follows with $\gamma^* = 2C_1/C_2$. ∎

Lemma 5.5 *For any $\gamma > 0$ and $\delta > 0$, the function F_δ^γ possesses a fixed point.*

Proof: By Lemma 5.2, r_μ is a continuous function of μ. Since μ occupies a compact set, so does the set $R = \{r_\mu | \mu \in \gamma\}$. Let $\overline{R} = \max\{\|r\| | r \in R\}$.

We only have to prove that a fixed point exists for a particular $\gamma > 0$, since, by Lemma 5.3 this fixed point is also a fixed point for all other positive values of γ.

Set $\gamma > 0$ such that there is a $\beta \in (0,1)$ with

$$\|F_\mu^\gamma(r) - r_\mu\|_2 \leq \beta \|r - r_\mu\|_2,$$

for all μ. (Existence of such a γ is ensured by Lemma 5.4.) We then have

$$\begin{aligned}
\|F_\delta^\gamma(r)\|_2 &\leq \|F_\delta^\gamma(r) - r_{\mu_r^\delta}\|_2 + \|r_{\mu_r^\delta}\|_2 \\
&\leq \beta \|r - r_{\mu_r^\delta}\|_2 + \overline{R} \\
&\leq \beta \|r\|_2 + (1+\beta)\overline{R}.
\end{aligned}$$

It follows that $\mathcal{C} = \left\{ r \middle| \|r\|_2 \leq (1+\beta)\overline{R}/(1-\beta) \right\}$, is closed under F_δ^γ. The result is then a consequence of Brouwer's fixed point theorem. ∎

Theorem 5.1 follows from Lemmas 5.3 and 5.5.

5.1 Existence of Fixed Points for H

Note that by replacing F_δ^γ with F^γ (defined in the same way as F_δ^γ with T replacing T_δ and $\mu_r = \mu_{\Phi r}$ replacing μ_r^δ), all steps in the proof of Lemma 5.5 remain valid except for the application of Brouwer's fixed point theorem, which can no longer be applied because F^γ may not be continuous because the greedy policy μ_r and the invariant distribution π_{μ_r} are not continuous in r. Nevertheless, Theorem 5.1 allows us to identify a sufficient condition for H to have a fixed point.

Let $\mathcal{V}_\delta$ be the set of fixed points of F_δ^γ, and let $\mathcal{P}$ be the set of vectors r such that more than one policy is greedy with respect to Φr. It is easy to show that $\mathcal{P}$ is closed. Finally, let $\mathcal{Q}_\epsilon = \{r | \|r - \hat{r}\| \geq \epsilon \text{ for all } \hat{r} \in \mathcal{P}\}$.

Theorem 5.2 *Suppose that there is $\epsilon > 0$, a decreasing sequence δ_k converging to 0, and a sequence $r_k \in \mathcal{V}_{\delta_k} \cap \mathcal{Q}_\epsilon$. Then, there exists r^* such that $\Phi r^* = H \Phi r^*$.*

Proof: First, let

$$\begin{aligned}
a_r(x) &= \operatorname*{argmax}_{a \in \mathcal{A}_x} \{g_a(x) + P_a(x) \Phi r\} \text{ and} \\
\Delta_\delta(r, x, a) &= g_{a_r(x)}(x) + \alpha P_{a_r(x)}(x) \Phi r - g_a(x) - \\
&\qquad - \alpha P_a(x) \Phi r.
\end{aligned}$$

Then $\inf_{r \in \mathcal{Q}_\epsilon, x, a \neq a_r(x)} \Delta_\delta(r, x, a) = \bar{\Delta} > 0$, and for all $r \in \mathcal{Q}_\epsilon$,

$$\mu_r^\delta(x, a_r(x)) \geq \left[1 + (m-1)\exp\left[-\bar{\Delta}/\delta\right]\right]^{-1},$$

where m is the maximum number of actions per state. Let $s(r) = \langle \phi_k, T_{\mu_r} \Phi r - \Phi r \rangle_{\mu_r}$. For small enough δ,

$$\|\mu_r - \mu_r^\delta\|_\infty \leq (m-1)\exp\left[-\bar{\Delta}/\delta\right],$$

thus μ_r^δ converges uniformly to μ_r in $\mathcal{Q}_\epsilon$. Now

$$\begin{aligned}
\|s(r_k)\|_\infty &= \|s(r_k) - s_\delta(r_k)\|_\infty \\
&= \Big\| \Phi' \Big[D_{\mu_{r_k}} \left(g_{\mu_{r_k}} + (\alpha P_{\mu_{r_k}} - I) \Phi r_k \right) - \\
&\qquad - D_{\mu_{r_k}^\delta} \left(g_{\mu_{r_k}^\delta} + (\alpha P_{\mu_{r_k}^\delta} - I) \Phi r_k \right) \Big] \Big\|_\infty \\
&\leq \|\Phi' D_{\mu_{r_k}} g_{\mu_{r_k}} - \Phi' D_{\mu_{r_k}^\delta} g_{\mu_{r_k}^\delta}\|_\infty + \\
&\quad + \max_i \Big\| \Phi' \Big[(\alpha P'_{\mu_{r_k}} - I) D_{\mu_{r_k}} - \\
&\qquad - (\alpha P'_{\mu_{r_k}^\delta} - I) D_{\mu_{r_k}^\delta} \Big) \Big] \phi_i \Big\|_2 \frac{(1+\beta)\overline{R}}{1-\beta}.
\end{aligned}$$

Since D_μ, g_μ and P_μ are continuous functions of μ and μ_r^δ converges to μ_r, $s(r_k)$ converges to 0. Hence, H has a fixed point. ∎

After proving that H_δ possesses at least one fixed point, it would be desirable to show that it is unique. Unfortunately, it can be shown by counterexamples that this is not always the case.

6 TD and its Stationary Points

The version of approximate value iteration that we have presented is related to and motivated by TD. The latter is a stochastic algorithm that adapts approximation weights r during simulation of the underlying Markov decision process. In this section, we will describe a version of the algorithm known as TD(0) and discuss how its stationary points coincide with fixed points of approximate value iteration.

Application of TD(0) entails simulating a single endless trajectory x_t of the Markov decision process under consideration. The weight vector is updated upon each transition, generating a sequence r_t. Given the state x_t and decision a_t at time t, if the next state is x_{t+1}, the weight vector r_t is updated according to

$$\begin{aligned} r_{t+1} &= r_t + \gamma_t \phi(x_t) \left(g_{a_t}(x_t) + \alpha(\Phi r_t)(x_{t+1}) - \right. \\ &\quad \left. -(\Phi r_t)(x_t)\right). \end{aligned}$$

But how is the decision a_t selected? One simple approach is to make "greedy decisions" with respect to the current estimate of the value function Φr_t. Such decisions are optimal if $\Phi r_t = J^*$. The hope is that, even though weights may initially lead to inaccurate approximations of J^* and poor decisions, as the simulation progresses, weights will converge to those that generate accurate approximations and near–optimal decisions. Unfortunately, the use of greedy decisions in TD(0) has appeared to perform poorly in practice (e.g., see [11]). Experiments point to the importance of "exploration;" i.e., randomization of the policy. One approach to exploration, which is connected to the variant of approximate value iteration studied in previous sections, selects decisions by letting $a_t = a$ with probability $\mu^{\delta}_{r_t}(x_t, a)$, for each $a \in \mathcal{A}_x$. Results from stochastic approximation theory (e.g., see [2]), show that if step sizes γ_t diminish at an appropriate rate, the process followed by r_t asymptotically approximates an ordinary differential equation

$$\dot{r} = \Phi' D_{\mu^{\delta}_r}(g_{\mu^{\delta}_r} + \alpha P_{\mu^{\delta}_r} \Phi r - \Phi r).$$

Intuitively, this ordinary differential equation drives r in the expected direction that would be taken by the stochastic algorithm. In particular,

$$\begin{aligned} &\Phi' D_{\mu^{\delta}_r}(g_{\mu^{\delta}_r} + \alpha P_{\mu^{\delta}_r} \Phi r - \Phi r) \\ &= \sum_{x \in S} \pi_{\mu^{\delta}_r}(x) \phi(x) \left(g_{\mu^{\delta}_r}(x) + \alpha(P_{\mu^{\delta}_r} \Phi r)(x) - (\Phi r)(x)\right). \end{aligned}$$

It is easy to show that a vector r is a stationary point of this ordinary differential equation if and only if it is a fixed point of our version of approximate value iteration. It follows that TD(0) with this form of exploration possesses stationary points. Note that, if greedy decisions were employed, the expected update direction when the weight vector is r would be zero if and only if Φr were a fixed point of H, which was shown in Example 3.2 not to necessarily have fixed points.

In this paper, we have established existence of fixed/stationary points for appropriate versions of approximate value iteration and TD. This is just one basic property, and a number of important questions remain open. Let us close by mentioning two: Do the proposed versions of approximate value iteration and/or TD converge? How well do fixed points of approximate value iteration approximate the optimal value function?

Acknowledgements

The authors would like to thank J.N. Tsitsiklis for comments on this paper. This research was supported in part by a Frederick E. Terman Award.

References

[1] Bellman, R. and Dreyfus, S., *Functional Approximations and Dynamic Programming*, Mathematical Tables and Other Aids to Computation, Vol. 13, pp 247-251, 1959.

[2] Benveniste, A., Métivier, M. and Priouret, P., *Adaptive Algorithms and Stochastic Approximation*, Springer-Verlag, Berlin, 1990.

[3] Bertsekas, D. P. and Tsitsiklis, J.N., *Neuro-Dynamic Programming*, Athena Scientific, 1995.

[4] Dayan, P.D., *The Convergence of TD(λ) for General λ*, Machine Learning, Vol. 8, pp. 341-362, 1992.

[5] Gallager, R.G., *Discrete Stochastic Processes*, Kluwer Academic Publishers, Boston, MA, 1996.

[6] Gurvits, L., Lin, L.J. and Hanson, S.J., *Incremental Learning of Evaluation Functions for Absorbing Markov Chains: New Methods and Theorems*, preprint, 1994.

[7] Pineda, F., *Mean-Field Analysis for Batched TD(λ)*, Neural Computation, pp. 1403–1419, 1997.

[8] Sutton, R.S., *Learning to Predict by the Method of Temporal Differences*, Machine Learning, Vol. 3, pp. 9-44, 1988.

[9] Tsitsiklis, J.N. and Van Roy, B., *An Analysis of Temporal-Difference Learning with Function Approximation*, IEEE Transactions on Automatic Control, Vol. 42, pp 674-690, 1997.

[10] Van Roy, B., *Learning and Value Function Approximation in Complex Decision Processes*, Ph.D. dissertation, MIT, 1998.

[11] Van Roy, B., Bertsekas, D.P., Lee, Y. and Tsitsiklis, J.N., *A Neuro-Dynamic Programming Approach to Retailer Inventory Management*, Proceedings of the IEEE Conference on Decision and Control, 1997.