# On the Existence of Fixed Points for Approximate Value Iteration and Temporal–Difference Learning[1]

D.P. de Farias[2] and B. Van Roy[3]

[1] This research was supported in part by a Frederick E. Terman Award.

[2] Graduate student, Department of Management Science and Engineering, Stanford University, Stanford, California.

[3] Assistant Professor, Department of Management Science and Engineering, Stanford University, Stanford, California.

## Abstract

Approximate value iteration is a simple algorithm that combats the curse of dimensionality in dynamic programs by approximating iterates of the classical value iteration algorithm in a spirit reminiscent of statistical regression. Each iteration of this algorithm can be viewed as an application of a modified dynamic programming operator to the current iterate. The hope is that the iterates converge to a fixed point of this operator, which will then serve as a useful approximation of the optimal value function. In this paper, 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. We then propose a variant of approximate value iteration for which the associated operator *is* guaranteed to possess at least one fixed point. This variant is motivated by studies of temporal–difference learning (TD), and existence of fixed points here implies existence of stationary points for the ordinary differential equation approximated by a version of TD that incorporates "exploration."

Key Words: Dynamic programming, neuro–dynamic programming, reinforcement learning, temporal–difference learning, value iteration.

# 1 Introduction

Value iteration offers a simple approach to computing optimal value functions and policies for finite–state discounted dynamic programs. The algorithm can be described compactly in terms of the "dynamic programming operator" $T$. In particular, value iteration 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 $J^*$, which is the unique fixed point of $T$ and can be used to generate an optimal policy.

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 approximation method – dating all the way back to Ref. [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 a spirit reminiscent of statistical regression. In rough terms, iterates $\tilde{J}_k$ are generated according to $\tilde{J}_{k+1} = \Pi T \tilde{J}_k$, where $\Pi$ is a projection operator that produces a function that is 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 concerning approximate value iteration is whether the composition $\Pi T$ 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 $\Pi T$ 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(Refs. [2]-[7]). As will be discussed in our closing section, 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 $x$th 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. In particular, when controlled by a policy $\mu$, the system evolves as a Markov chain with transition probabilities

$$P_\mu(x, y) = \sum_{a \in \mathcal{A}_x} \mu(x, a) P_a(x, y).$$

For shorthand notation, let

$$g_\mu(x) = \sum_{a \in \mathcal{A}_x} \mu(x, a) g_a(x).$$

A policy $\mu$ is deterministic if for each $x \in S$, there is an action $a \in \mathcal{A}_x$ such that $\mu(x, a) = 1$. A policy $\mu$ 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 $x \in \mathcal{S}$, or equivalently, it attains the supremum of

$$\sum_{k=0}^{\infty} (\alpha^k P_\mu^k) g_\mu,$$

for all entries.

We assume that $P_\mu$ is irreducible and aperiodic for every $\mu$. Hence, each $P_\mu$ possesses a unique invariant distribution $\pi_\mu$ with $\pi_\mu(x) > 0$ for all $x$. Note that the invariant distribution is the left eigenvector associated with the unit eigenvalue of $P_\mu$ with entries adding to one.

The optimal value function $J^*$ uniquely solves Bellman's equation:

$$J = \max_\mu \left\{g_\mu + \alpha P_\mu J\right\},$$

where the maximization is pointwise. Alternatively, defining the dynamic programming operator $T$ by

$$TJ = \max_\mu \left\{g_\mu + \alpha P_\mu J\right\},$$

the value function can be characterized as the unique fixed point of $T$. For each policy $\mu$, we define an additional operator $T_\mu$ by

$$T_\mu J = g_\mu + \alpha P_\mu J.$$

It is well–known that for any $J$, there is a deterministic policy $\mu$ such that $TJ = T_\mu J$. We call such a $\mu$ a *greedy policy* with respect to $J$. It is also well–known that a policy $\mu^*$ 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 a sequence according to $J_{k+1} = TJ_k$, initialized with some function $J_0$. 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 compactly represented approximations rather than functions over the state space. In particular, given a preselected collection $\phi_1, \ldots, \phi_K$ of basis functions, the algorithm generates approximations $\tilde{J}_k$ to each iterate $J_k$, where

$$\tilde{J}_k = \sum_{i=1}^{K} r_k(i)\phi_i,$$

for some weight vector $r_k \in \Re^K$. Defining a $|\mathcal{S}| \times K$ matrix

$$\Phi = \begin{bmatrix} | & & | \\ \phi_1 & \cdots & \phi_K \\ | & & | \end{bmatrix},$$

we have $\tilde{J}_k = \Phi r_k$. Also, let $\phi(x) = (\phi_1(x), \ldots, \phi_K(x))'$ 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 $\Pi$ that projects onto the span of $\phi_1, \ldots, \phi_K$ with respect to the standard Euclidean norm, i.e.,

$$\Pi J = \operatorname*{argmin}_{\Phi r} \|J - \Phi r\|_2.$$

The algorithm then generates iterates according to $\tilde{J}_{k+1} = \Pi T \tilde{J}_k$. Hence, the operator $T$ is effectively approximated by $\Pi T$.

A slightly more sophisticated version of approximate value iteration projects with respect to a weighted Euclidean norm. In particular, given a $|\mathcal{S}|$–dimensional vector $\pi$ with positive components, we can define a norm by

$$\|J\|_\pi = \left( \sum_{x \in \mathcal{S}} \pi(x) J^2(x) \right)^{1/2},$$

and the associated projection is given by

$$\Pi J = \operatorname*{argmin}_{\Phi r} \|J - \Phi r\|_\pi.$$

As before, approximate value iteration would generate iterates according to $\tilde{J}_{k+1} = \Pi T \tilde{J}_k$. One possible 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, a more general form of approximate value iteration 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 $\Pi_J$, approximate value iteration could generate iterates according to $\tilde{J}_{k+1} = \Pi_{\tilde{J}_k} T \tilde{J}_k$.

# 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, the hope that iterates $\tilde{J}_k$ will converge to a fixed point can not be met. We provide in this section two examples that illustrate difficulties that arise and motivate the variant of approximate value iteration that we will propose in the next section.

Our first example makes use of a projection operator $\Pi$ that simply projects with respect to the standard Euclidean norm. In this example, the composition $\Pi T$ does not possess a fixed point, and furthermore, approximate value iteration leads to an unbounded sequence of iterates.

**Example 3.1** *Consider a Markov chain with states 1 and 2 and only one policy. 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 we want to find an approximate value function in the subspace spanned by $\Phi = [1\ 2]'$. Then if $\Pi$ is the projection with respect to the Euclidean norm, if $\Phi r = \Pi T \Phi r$, the parameter vector $r$ must solve*

$$\begin{aligned} r &= (\Phi'\Phi)^{-1}\Phi'(g + \alpha P \Phi r) \\ &= 1 + r. \end{aligned}$$

*Hence, $\Pi T$ 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 Refs. [5, 8] 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 $\pi$ be the invariant distribution. Then, $\|P\|_\pi \leq 1$.*

An immediate consequence of this lemma is that, for a problem with only one policy (i.e., $|\mathcal{A}_x| = 1$ for all $x$), the operator $T$ is a contraction; in particular,

$$\|TJ - T\overline{J}\|_\pi \leq \alpha \|J - \overline{J}\|_\pi,$$

for any $J$ and $\overline{J}$, where $\pi$ is the invariant distribution associated with the single policy. Furthermore, since a projection matrix $\Pi$ that projects with respect to $\|\cdot\|_\pi$ is nonexpansive with respect to $\|\cdot\|_\pi$, we have

$$\|\Pi TJ - \Pi T\overline{J}\|_\pi \leq \alpha \|J - \overline{J}\|_\pi,$$

for any $J$ and $\overline{J}$. In other words, $\Pi T$ is a contraction. It follows that $\Pi T$ has a unique fixed point and that value iteration converges to this fixed point.

The above discussion identifies a version of value iteration that converges to a fixed point. Unfortunately, this only applies to problems with a single policy. Can the essential idea be generalized to problems with multiple policies? In such cases, there are usually multiple invariant distributions to take into account, each associated with a different policy. One approach to dealing with this issue involves projecting with respect to a Euclidean norm weighted by the invariant distribution associated with a policy that is greedy with respect to the current iterate $\tilde{J}_k$. In particular, let $\mu_J$ be a greedy policy with respect to $J$, and let for any $\mu$, let $\Pi_\mu$ be a projection matrix that projects onto the span of $\phi_1, \ldots, \phi_K$ with respect to $\|\cdot\|_{\pi_\mu}$. A version of approximate value iteration can be defined by $\tilde{J}_{k+1} = \Pi_{\mu_J} TJ$. We define the operators $H$ and $H_\mu$ by

$$HJ = \Pi_{\mu_J} TJ \quad \text{and} \quad H_\mu J = \Pi_\mu T_\mu J.$$

To make this definition unambiguous, we need to establish what policy $\mu_J$ must be used if there is more than one greedy policy – in this case, we let $\mu_J$ be the randomized policy that takes each greedy action with the same probability.

The question that arises is whether $H$ possesses a fixed point. In the case of a single policy the question reduces to one that we have already addressed, and the answer is affirmative. The following example, adapted from Ref. [7], shows that $H$ does not necessarily possess 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*

$$g_1 = g_2 = \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix}, P_1 = \begin{bmatrix} 0.2 & 0 & 0.8 \\ 0.4 & 0.6 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad P_2 = \begin{bmatrix} 0.2 & 0 & 0.8 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

*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 $\mu_J$: $\mu_1$ (with $P_{\mu_1} = P_1$), $\mu_2$ (with $P_{\mu_2} = P_2$), and $\mu_3$ (with $P_{\mu_3} = (P_1 + P_2)/2$).*

*A function $J$ is a fixed point of $H$ if and only if $\Pi_{\mu_J} TJ = J$, or equivalently, $\Pi_{\mu_J} T_{\mu_J} J = J$. Based on arguments made earlier regarding approximate value iteration for problems with a single policy, 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.*

Based on the above example, one might speculate 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 show in the sequel, incorporating "exploration" (i.e., randomizing 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 $\delta$–greedy policies, which effectively incorporate exploration into a greedy policy. Formally, for any $\delta > 0$, we define a $\delta$–greedy policy $\mu_J^\delta$ with respect to $J$ by

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

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

$$T_\delta J = T_{\mu_J^\delta} J.$$

Note that $T_\delta$ is continuous.

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

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

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

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

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

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

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

**q.e.d.**

The following lemma establishes that $T_\delta$ approximates $T$ uniformly as $\delta \downarrow 0$ and follows as an immediate consequence of the previous lemma.

**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 $\delta > 0$, $T_\delta$ 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 $T_\delta$, and since $T_\delta$ is continuous, Brouwer's fixed point theorem guarantees existence of a fixed point. **q.e.d.**

Let us introduce the notion of a *quasi-contraction*, which will help us study the operator $T_\delta$ and its fixed points.

**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 $\delta > 0$, $T_\delta$ is a quasi-contraction.*

**Proof:** For any $J_1$ and $J_2$,

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

and since we know from Lemma 4.3 that $T_\delta$ has a fixed point, $T_\delta$ is indeed a quasi-contraction. **q.e.d.**

The following lemma establishes that, for small $\delta$, fixed points of $T_\delta$ 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 $T_\delta$. 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 $\|J - J^*\|_\infty = O(\delta)$. Note that the $O(\delta)$ term in the final inequality is uniformly bounded over $J$ by Lemma 4.2. **q.e.d.**

Lemma 4.4 bears some important implications on $T_\delta$. First, note that all fixed points of $T_\delta$ lie within a ball of radius $O(\delta)$. Furthermore, for $J$ outside this circle, $T_\delta$ 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 $T_\delta$, we can define a new version of approximate value iteration, which updates weights according to

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

Alternatively, defining an operator $H_\delta$ by

$$H_\delta J = \Pi_{\mu^\delta_J} T_\delta J,$$

we have $\tilde{J}_{k+1} = H_\delta \tilde{J}_k$. The following theorem, which is the main result of this section, establishes that, unlike $H$, $H_\delta$ always possesses a fixed point.

**Theorem 5.1** *For any $\delta > 0$, $H_\delta$ 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 $\Pi^\delta_r$ to refer to $\Pi_{\mu^\delta_{\Phi r}}$ and $\mu^\delta_r$ to refer to $\mu^\delta_{\Phi r}$.

## 5.1 Preliminary Lemmas

We begin by establishing continuity of certain functions that are important to our analysis.

**Lemma 5.1** *The invariant distribution $\pi_\mu$ is a continuous function of $\mu$.*

**Proof:** Since $P_\mu$ is irreducible, all but one of its eigenvalues are strictly inside the unit circle. The remaining eigenvalue is 1 and corresponds to a left eigenvector of $\pi_\mu$ Ref. [9]. It follows that the matrix

$$\begin{bmatrix} P'_\mu - I \\ e' \end{bmatrix}$$

has full column rank. As an invariant distribution, $\pi_\mu$ uniquely satisfies

$$\begin{bmatrix} P'_\mu - I \\ e' \end{bmatrix} \pi_\mu = \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

and therefore,

$$\pi_\mu = \begin{bmatrix} P'_\mu - I \\ e' \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Since $P_\mu$ is a continuous function of $\mu$, so is $\pi_\mu$. Note that for a full column rank matrix $A$, not necessarily square, we let $A^{-1} = (A'A)^{-1}A'$. **q.e.d.**

As discussed earlier, for each policy $\mu$, there exists a unique vector $r_\mu$ 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 $\mu$.

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

**Proof:** The equation $\Phi r_\mu = H_\mu \Phi r_\mu$ can be rewritten as

$$\Phi r_\mu = \Pi_\mu \left( g_\mu + \alpha P_\mu \Phi r_\mu \right),$$

or, via rearranging, as

$$(I - \alpha \Pi_\mu P_\mu) \Phi r_\mu = \Pi_\mu g_\mu.$$

¿From Ref. [8], we know that $\|\Pi_\mu P_\mu\|_\mu \leq 1$, and therefore all eigenvalues of $\Pi_\mu P_\mu$ are in the unit circle (possibly on the boundary). It follows that $I - \alpha \Pi_\mu P_\mu$ is invertible, and we have

$$\Phi r_\mu = (I - \alpha \Pi_\mu P_\mu)^{-1} \Pi_\mu g_\mu.$$

and since $\Phi$ has full column rank,

$$r_\mu = \Phi^{-1}(I - \alpha\Pi_\mu P_\mu)^{-1}\Pi_\mu g_\mu.$$

By Lemma 5.1, $D_\mu = \mathrm{diag}(\pi_\mu)$ is a continuous function of $\mu$, and therefore, $\Pi_\mu = \Phi(\Phi' D_\mu \Phi)^{-1}\Phi' D_\mu$ is also a continuous function of $\mu$. Continuity of $r_\mu$ follows. **q.e.d.**

## 5.2 Main Analysis

For any policy $\mu$, let us define

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

where $D_\mu = \mathrm{diag}(\pi_\mu)$ for any policy $\mu$. Furthermore, we define functions $F^\gamma_\mu : \Re^K \mapsto \Re^K$ and $F^\gamma_\delta : \Re^K \mapsto \Re^K$ by

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

The following lemma relates fixed points of $F^\gamma_\mu$ and $F^\gamma_\delta$ to those of $H_\mu$ and $H_\delta$.

**Lemma 5.3** *For any $\delta > 0$ and $\gamma > 0$, a vector $r$ is a fixed point of $F^\gamma_\mu$ ($F^{\delta,\gamma}$) if and only if $\Phi r$ is a fixed point of $H_\mu$ ($H_\delta$).*

**Proof:** Let $r$ be a fixed point of $F^\gamma_\mu$. Then,

$$\begin{aligned}
s_\mu(r) &= 0 \\
\Phi' D_\mu \Phi r &= \Phi' D_\mu(g_\mu + \alpha P_\mu \Phi r) \\
\Phi(\Phi' D_\mu \Phi)^{-1}\Phi' D_\mu \Phi r &= \Phi(\Phi' D_\mu \Phi)^{-1}\Phi' D_\mu(g_\mu + \alpha P_\mu \Phi r) \\
\Phi r &= \Pi_\mu T_\mu \Phi r,
\end{aligned}$$

and $\Phi r$ is a fixed point of $H_\mu$. Reversing the steps, it can be shown that if $\Phi r$ is a fixed point of $H_\mu$, $r$ is a fixed point of $F^\gamma_\mu$.

An entirely analogous argument establishes that $r$ is a fixed point of $F^\gamma_\delta$ if and only if $\Phi r$ is a fixed point of $H_\delta$. **q.e.d.**

The next lemma establishes that, for sufficiently small $\gamma$, $F^\gamma_\mu$ is a pseudo–contraction.

**Lemma 5.4** *There exists a constant $\gamma^* > 0$ such that for all $\mu$ 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 $\mu$,

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

and

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

Since $D_\mu$ is positive definite for all $\mu$ and the set of all randomized policies is compact, it follows that there exists a constant $C_1 > 0$ independent of $\mu$ such that

$$(r - r_\mu)' s_\mu(r) \leq -C_1 \|r - r_\mu\|^2_2.$$

Note that

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

and it follows that there exists a constant $C_2 > 0$ independent of $\mu$ such that

$$\|s_\mu(r)\|_2^2 \leq C_2 \|r - r_\mu\|_2^2.$$

Making use of the inequalities we have established, we have

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

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

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

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

Note that we only have to establish 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 $\gamma$.

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 $\mu$. (Existence of such a $\gamma$ 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 the set

$$\mathcal{C} = \left\{ r \middle| \|r\|_2 \leq \frac{(1+\beta)\overline{R}}{1-\beta} \right\},$$

is closed under $F_\delta^\gamma$. The result is then a consequence of Brouwer's fixed point theorem. **q.e.d.**

Theorem 5.1 follows from Lemmas 5.3 and 5.5.

## 5.3 Existence of Fixed Points for $H$

Note that by replacing $F_\delta^\gamma$ with $F^\gamma$ (defined in the same way as $F_\delta^\gamma$ with $T$ replacing $T_\delta$ and $\mu_r = \mu_{\Phi r}$ replacing $\mu_r^\delta$), 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 $F^\gamma$ may not be continuous because the greedy policy $\mu_r$ and the invariant distribution $\pi_{\mu_r}$ are not continuous in $r$. Nevertheless, Theorem 5.1 allows us to identify a sufficient condition for $H$ to have a fixed point, as will be established in the upcoming theorem.

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

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

**Proof:** First, let

$$a_r(x) = \operatorname*{argmax}_{a \in \mathcal{A}_x} \left\{ g_a(x) + P_a(x) \Phi r \right\},$$

and

$$\Delta_\delta(r, x, a) = g_{a_r(x)}(x) + \alpha P_{a_r(x)}(x) \Phi r - g_a(x) - \alpha P_a(x) \Phi r.$$

Then,

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

and for all $r \in \mathcal{Q}_\epsilon$,

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

where $m$ is the maximum number of actions per state. Let

$$s(r) = \langle \phi_k, T_{\mu_r} \Phi r - \Phi r \rangle_{\mu_r}.$$

Note that for small enough $\delta$,

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

and therefore, $\mu_r^\delta$ converges uniformly to $\mu_r$ in $\mathcal{Q}_\epsilon$. Now

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

Since $D_\mu$, $g_\mu$ and $P_\mu$ are continuous functions of $\mu$ and $\mu_r^\delta$ converges to $\mu_r$, $s(r_k)$ converges to 0. Hence, $H$ has a fixed point. **q.e.d.**

# 6 Multiplicity of Fixed Points

It was established in the previous section that $H_\delta$ possesses at least one fixed point. It would be convenient if the fixed point were known to be unique, but unfortunately, this is not always the case, as demonstrated by the following example.

**Example 6.1** *Consider again a controlled Markov chain with three states and two deterministic policies 1 and 2. Let the transistion matrices be the same as in Example 3.2 and the rewards be*

$$g_1 = g_2 = \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix} \tag{1}$$

*Let* $\alpha = 0.99$ *and* $\Phi = [0\ 1\ 2]'$. *Then policies 1 and 2 have approximate value functions with parameters* $r_1 = 0.1647$ *and* $r_2 = -0.3311$. *The optimal policies for the one-step problem with final rewards* $\Phi r_1$ *and* $\Phi r_2$ *are, respectively, 1 and 2, hence* $H$ *has two fixed points. We are also able to find that for* $\delta = 0.001$, $H_\delta$ *has a fixed point between -0.3311264 and -0.3311256 and another fixed point between 0.1647443 and 0.1647449.*

# 7 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. We depart, though, from the degree of rigor maintained in previous sections and only present heuristic arguments.

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

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

But how is the decision $a_t$ selected? One simple approach that has been proposed makes "greedy decisions;" that is,

$$a_t = \operatorname*{argmax}_{a \in \mathcal{A}_x} \left( g_a(x) + \alpha(P_a \Phi r_t)(x) \right).$$

Note that such decisions are optimal if $\Phi r_t = J^*$. The hope is that, 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 Ref. [10]). 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 $a \in \mathcal{A}_x$. Using results from stochastic approximation theory (e.g., see Ref. [11]), one can show that if step sizes $\gamma_t$ 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, where the expectation is taken over the steady state distribution of $\mu_r^\gamma$. In particular,

$$\Phi' D_{\mu_r^\delta}(g_{\mu_r^\delta} + \alpha P_{\mu_r^\delta}\Phi r - \Phi r) = \sum_{x \in S} \pi_{\mu_r^\delta}(x)\phi(x)\left(g_{\mu_r^\delta}(x) + \alpha(P_{\mu_r^\delta}\Phi r)(x) - (\Phi r)(x)\right).$$

A vector $r$ is a stationary point of this ordinary differential equation if and only if

$$\begin{aligned}
\Phi' D_{\mu_r^\delta}\Phi r &= \Phi' D_{\mu_r^\delta}(g_{\mu_r^\delta} + \alpha P_{\mu_r^\delta}\Phi r) \\
\Phi(\Phi' D_{\mu_r^\delta}\Phi)^{-1}\Phi' D_{\mu_r^\delta}\Phi r &= \Phi(\Phi' D_{\mu_r^\delta}\Phi)^{-1}\Phi' D_{\mu_r^\delta}(g_{\mu_r^\delta} + \alpha P_{\mu_r^\delta}\Phi r) \\
\Phi r &= \Pi_{\mu_r^\delta} T_\delta \Phi r.
\end{aligned}$$

Hence, stationary points coincide with fixed points 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 are employed, the expected update direction when the weight vector is $r$ is given by

$$\Phi' D_{\mu_r}(g_{\mu_r} + \alpha P_{\mu_r}\Phi r - \Phi r).$$

For this quantity to be zero, we must have

$$\Phi r = \Pi_{\mu_r} T \Phi r,$$

i.e., $\Phi r$ must be a fixed point of the version of approximate value iteration considered in Example 2. As illustrated in that example, the associated operation need not possess a fixed point.

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:

1. Do the proposed versions of approximate value iteration and/or TD converge?
2. How well do fixed points of approximate value iteration approximate the optimal value function?

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

[3] GURVITS, L., LIN, L.J. and HANSON, S.J., *Incremental Learning of Evaluation Functions for Absorbing Markov Chains: New Methods and Theorems*, preprint, 1994.

[4] PINEDA, F., *Mean-Field Analysis for Batched TD($\lambda$)*, Neural Computation, pp. 1403–1419, 1997.

[5] 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.

[6] DAYAN, P.D., *The Convergence of TD($\lambda$) for General* $\lambda$, Machine Learning, Vol. 8, pp. 341-362, 1992.

[7] BERTSEKAS, D. P. and TSITSIKLIS, J.N., *Neuro-Dynamic Programming*, Athena Scientific, 1995.

[8] VAN ROY, B., *Learning and Value Function Approximation in Complex Decision Processes*, Ph.D. dissertation, MIT, 1998.

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

[10] 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.

[11] BENVENISTE, A., MÉTIVIER, M. and PRIOURET, P., *Adaptive Algorithms and Stochastic Approximation*, Springer-Verlag, Berlin, 1990.