Tight Performance Bounds on Greedy Policies Based on Imperfect Value Functions *

Ronald J. Williams
College of Computer Science
Northeastern University
Boston, MA 02115
rjw@ccs.neu.edu

Leemon C. Baird, III
Wright Laboratory
Wright-Patterson Air Force Base, OH 45433-6543
bairdlc@wL.wpafb.af.mil

Northeastern University
College of Computer Science
Technical Report NU-CCS-93-14

November 24, 1993

Abstract

Consider a given value function on states of a Markov decision problem, as might result from applying a reinforcement learning algorithm. Unless this value function equals the corresponding optimal value function, at some states there will be a discrepancy, which is natural to call the Bellman residual, between what the value function specifies at that state and what is obtained by a one-step lookahead along the seemingly best action at that state using the given value function to evaluate all succeeding states. This paper derives a tight bound on how far from optimal the discounted return for a greedy policy based on the given value function will be as a function of the maximum norm magnitude of this Bellman residual. A corresponding result is also obtained for value functions defined on state-action pairs, as are used in Q-learning. One significant application of these results is to problems where a function approximator is used to learn a value function, with training of the approximator based on trying to minimize the Bellman residual across states or state-action pairs. When

*This work was supported by Grant IRI-8921275 from the National Science Foundation and by the U. S. Air Force.

control is based on the use of the resulting value function, this result provides a link between how well the objectives of function approximator training are met and the quality of the resulting control.

1 Introduction

This paper examines the question of how far from optimal the discounted return arising from a policy can be, expressed as a function of the kind of value function error typically used in reinforcement learning applications. The dependent variable in this functional relationship is the difference between the actual return and the optimal return, and the independent variable is what we call the *Bellman equation error*. The primary significance of this quantity is that it corresponds very naturally to what most reinforcement learning methods actually try to minimize. Thus the results presented here provide a direct link between the objectives of such algorithms and the quality of the resulting control, under the realistic assumption that perfect learning (meaning zero Bellman equation error) does not occur.

Singh and Yee (to appear) have also derived bounds of this type, but there are some important differences between their results and those presented here. The main results they derive use as the independent variable the max-norm distance between the value function and the optimal value function rather than the Bellman equation error. Combining these results with a standard result then leads to additional results of the type considered here, but the resulting bounds turn out to be far less tight than those found through the more direct route taken here.

We derive two sets of bounds, one for value functions defined on states only, and another for value functions defined on state-action pairs, as used in the Q-learning algorithm (Watkins, 1989; Watkins & Dayan, 1992). Since Q-learning has become the most prominent reinforcement learning algorithm, the results presented in the section dealing with state-action pairs are probably of greatest interest. One important reason for including the corresponding results for state value functions, and for presenting these results first, is that the simplicity of the arguments is more apparent in this case. The arguments for the state-action value function case correspond very closely to those for the state value case, but the extra machinery required, most of it unfamiliar in standard dynamic programming theory and therefore in need of more detailed supporting arguments, tends to obscure the underlying simplicity. Thus the material on state value functions is included both for independent interest and to help pave the way for understanding the state-action value case.

2 Markov Decision Problem and Dynamic Programming

Here we give a brief overview of the fundamental notions of stochastic dynamic programming and introduce the mathematical notation to be used throughout this paper. A more detailed description, along with proofs of results we cite as part of standard dynamic programming theory, may be found in Bertsekas (1987).

We take as given a *Markov environment*, or *controlled Markov chain*, having a set of states X and a set of actions A. We assume that both X and A are finite. We let $f(x, a)$ denote the randomly determined successor of state x when action a is applied. The behavior of this random next-state function is determined by the transition probabilities $p^a_{xy} = Pr\{f(x,a) = y\}$ for $x, y \in X$ and $a \in A$. We also assume that associated with each choice of action a at each state x is a randomly determined immediate reward $r(x,a)$, with $R(x,a) = E\{r(x,a)\}$ denoting its expected value.

In general, a non-randomized policy is a function π assigning to each possible history of states and actions a choice of action to be used at the current time. Here we generally restrict attention to *stationary* policies, which select actions according to the current state only. Thus a stationary policy can be viewed as a function $\pi : X \rightarrow A$.

A *Markov decision problem* consists of a such a Markov environment together with a criterion function on policies, and the objective is to find a policy optimizing this criterion.

For any policy π, define the real-valued function V^π on states by

$$V^\pi(x) = E\left\{\sum_{t=0}^{\infty} \gamma^t r(x_t, a_t) \mid x_0 = x\right\}$$

where it is also given that $x_{t+1} = f(x_t, a_t)$ for all $t > 0$ and a_t is determined by the policy π for all $t \geq 0$. This quantity is called the *discounted return* for policy π at state x, and the discount parameter γ is assumed to lie in $[0, 1)$. We call any mapping from X into the real numbers a *state value function*, and we see that V^π is a special case of this notion.

Define a partial order relation on state value functions by $V \leq V'$ if and only if $V(x) \leq V'(x)$ for all $x \in X$. An *optimal* policy is one for which the return is maximal at each state. With V^* denoting the return from any optimal policy, it follows that $V^\pi \leq V^*$ for any policy π. Clearly V^* is unique if there are any optimal policies. A fundamental result from the theory of dynamic programming is that, under the conditions assumed here, there exist optimal stationary policies.

3 Results For State Value Functions

In this section we give definitions and derive results for the case when a state value function is used to determine a policy through the use of what amounts to a one-step lookahead.

3.1 Backup Operators

We define two types of backup operator on state value functions as follows. For any stationary policy π, $B^\pi V$ is that state value function assigning to state x the value

$$\begin{aligned} B^\pi V(x) &= E\left\{r(x, \pi(x)) + \gamma V(f(x, \pi(x)))\right\} \\ &= R(x, \pi(x)) + \gamma \sum_{y \in X} p^{\pi(x)}_{xy} V(y), \end{aligned}$$

while BV assigns to state x the value

$$\begin{aligned} BV(x) &= \max_{a \in A} E\left\{r(x,a) + \gamma V(f(x,a))\right\} \\ &= \max_{a \in A} \left[R(x,a) + \gamma \sum_{y \in X} p^a_{xy} V(y)\right]. \end{aligned}$$

Two standard results from the theory of dynamic programming are that, under the conditions assumed here, $V = V^\pi$ is the unique solution of the equation $V = B^\pi V$, and $V = V^*$ is the unique solution of the *Bellman equation* $V = BV$.

3.2 Greedy Policies

Given a state value function V, define a stationary policy π to be *greedy* for V if

$$\pi(x) = \arg\max_{a\in A}\left[R(x,a) + \gamma\sum_{y\in X} p_{xy}^a V(y)\right]$$

for all $x \in X$.

3.3 Maximum Norm Distance Measure

For the results presented here, distances between value functions are based on the maximum norm. For the case of state value functions, we thus define

$$\|V - V'\| = \|V - V'\|_\infty = \max_{x\in X}|V(x) - V'(x)|$$

for any two state value functions V and V'.

3.4 Bellman Residual

We single out for particular attention the *Bellman error magnitude* for a given state value function V, which is simply the max norm distance $\|BV - V\| = \max_x |BV(x) - V(x)|$ between the left-hand and right-hand sides of the Bellman equation. We also use the term *Bellman residual* or *Bellman equation error* for V to mean the state value function $BV - V$. For convenience, we will typically shorten these terms still further to *V-residual* and *V-error magnitude*.

The Bellman residual is significant for three reasons. First, since $V = V^*$ is the unique solution of the Bellman equation, it is zero if and only if $V = V^*$. Second, it is readily computable from the given value function, unlike a quantity like $V^* - V$, used in some other analyses of performance bounds on greedy policies (Singh & Yee, to appear), which requires knowledge of V^*. And most importantly for applications to learning, when training a function approximator to represent a value function on states (or state-action pairs, as considered below), the approach universally used is based on trying to minimize the individual temporal difference (TD) errors (Sutton, 1988), which are closely related to the Bellman residual. There is thus a very natural correspondence between what training a function approximator using TD errors tries to accomplish and what the Bellman residual measures.

3.5 Derivation of Performance Bounds

Here we derive the desired tight performance bounds, given in Theorem 3.2 and Corollary 3.1. We begin by stating without proof an easily derived standard contraction result from the theory of dynamic programming and then observing some simple consequences of it.

Lemma 3.1 *Given any two state value functions V and V' and stationary policy π,*

$$\|B^\pi V - B^\pi V'\| \leq \gamma \|V - V'\|$$

and

$$\|BV - BV'\| \leq \gamma \|V - V'\|$$

□

Proposition 3.1 *For any state value functions V and any policy π,*

$$\|V - V^\pi\| \leq \frac{\|V - B^\pi V\|}{1-\gamma}$$

and

$$\|V - V^*\| \leq \frac{\|V - BV\|}{1-\gamma}.$$

Proof. To prove the first inequality, we apply the triangle inequality, the first inequality of the previous lemma, and the fact that V^π is fixed by B^π to obtain

$$\|V - V^\pi\| \leq \|V - B^\pi V\| + \|B^\pi V - V^\pi\| \leq \|V - B^\pi V\| + \gamma \|V - V^\pi\|,$$

from which it follows that

$$\|V - V^\pi\| \leq \frac{\|V - B^\pi V\|}{1-\gamma}.$$

The second inequality follows in the same way, using instead the second inequality of the previous lemma and the fact that V^* is fixed by B^*. □

Now we are prepared to derive our first performance bound result. We will see later that this bound is not as tight as possible. Nevertheless we begin with this since the argument required for a tight bound is slightly more elaborate and since this simpler argument parallels that given later for state-action value functions.

Theorem 3.1 *Let V be a value function on X, and let π be a greedy policy for V. Let $\varepsilon = \|BV - V\|$ denote the Bellman error magnitude for V. Then*

$$V^\pi(x) \geq V^*(x) - \frac{2\varepsilon}{1-\gamma}$$

for any state x. Furthermore, if $V^* \leq V$, then

$$V^\pi(x) \geq V^*(x) - \frac{\varepsilon}{1-\gamma}$$

for any state x.

Proof. Note that π greedy for V implies $B^\pi V = BV$. We can then combine the two inequalities of the previous proposition with the triangle inequality to conclude that

$$\|V^* - V^\pi\| \leq \|V^* - V\| + \|V - V^\pi\| \leq \frac{2\|V - BV\|}{1-\gamma} = \frac{2\varepsilon}{1-\gamma}.$$

Since $V^\pi \leq V^*$, this implies

$$V^*(x) - V^\pi(x) = |V^*(x) - V^\pi(x)| \leq \frac{2\varepsilon}{1-\gamma}$$

for any state x, from which the first bound follows.

To establish the second bound, note that $V^* \leq V$ implies $V^\pi(x) \leq V^*(x) \leq V(x)$ for any state x and for any policy π, so when π is greedy for V,

$$\begin{aligned} V^*(x) - V^\pi(x) &\leq V(x) - V^\pi(x) \\ &\leq \frac{\|V - B^\pi V\|}{1-\gamma} \\ &= \frac{\|V - BV\|}{1-\gamma} \\ &= \frac{\varepsilon}{1-\gamma}. \end{aligned}$$

$\square$

The bounds given by Theorem 3.1 can be made tight by introducing an additional factor of γ, as we now show.

Theorem 3.2 *Let V be a value function on X, and let π be a greedy policy for V. Let $\varepsilon = \|BV - V\|$ denote the Bellman error magnitude for V. Then*

$$V^\pi(x) \geq V^*(x) - \frac{2\gamma\varepsilon}{1-\gamma}$$

for any state x. Furthermore, if $V^* \leq V$, then

$$V^\pi(x) \geq V^*(x) - \frac{\gamma\varepsilon}{1-\gamma}$$

for any state x. In addition, in each case there is an example where equality holds.

Proof. As before, $B^\pi V = BV$ since π is greedy for V. Therefore, for any state x, the first inequality of Proposition 3.1 implies that

$$V(x) \leq V^\pi(x) + \frac{\epsilon}{1-\gamma}, \tag{1}$$

while the second inequality of that proposition implies that

$$V^*(x) \leq V(x) + \frac{\epsilon}{1-\gamma}. \tag{2}$$

Now pick a state x. Let a be an optimal action at x and let $\pi(x) = b$. Since π is greedy for V, it follows that

$$R(x,a) + \gamma \sum_{y \in X} p_{xy}^a V(y) \leq R(x,b) + \gamma \sum_{y \in X} p_{xy}^b V(y). \tag{3}$$

We then use (2), (3), and (1) to conclude that

$$\begin{aligned}
V^*(x) &= R(x,a) + \gamma \sum_{y \in X} p_{xy}^a V^*(y) \\
&\leq R(x,a) + \gamma \sum_{y \in X} p_{xy}^a \left[V(y) + \frac{\epsilon}{1-\gamma} \right] \\
&= R(x,a) + \gamma \sum_{y \in X} p_{xy}^a V(y) + \frac{\gamma\varepsilon}{1-\gamma} \\
&\leq R(x,b) + \gamma \sum_{y \in X} p_{xy}^b V(y) + \frac{\gamma\varepsilon}{1-\gamma} \\
&\leq R(x,b) + \gamma \sum_{y \in X} p_{xy}^b \left[V^\pi(y) + \frac{\epsilon}{1-\gamma} \right] + \frac{\gamma\varepsilon}{1-\gamma} \\
&= R(x,b) + \gamma \sum_{y \in X} p_{xy}^b V^\pi(y) + \frac{2\gamma\varepsilon}{1-\gamma} \\
&= V^\pi(x) + \frac{2\gamma\varepsilon}{1-\gamma},
\end{aligned}$$

which proves the first inequality. The second inequality is proved in identical fashion, but with (2) replaced by the inequality $V^*(x) \leq V(x)$ for all x.

To see that the first bound cannot be made tighter in general, consider a Markov decision problem having two states 1 and 2 and two actions 1 and 2, where the effect of action i in either state is to cause a transition to state i for $i = 1$ or 2, with all immediate rewards being 0 except that $R(2,2) = 2$. Now consider the state value function V defined by

$$V(1) = V(2) = \frac{1}{1-\gamma}.$$

The stationary policy π with $\pi(1) = 1$ and $\pi(2) = 2$ is greedy for V, while the only optimal policy is to take action 2 in either state. The V-residual at state 1 is

$$BV(1) - V(1) = \gamma V(1) - V(1) = \frac{\gamma}{1-\gamma} - \frac{1}{1-\gamma} = -1,$$

while the V-residual at state 2 is

$$BV(2) - V(2) = 2 + \gamma V(2) - V(2) = 2 + \frac{\gamma}{1-\gamma} - \frac{1}{1-\gamma} = 1.$$

Thus the V-error magnitude is $\varepsilon = 1$. The optimal return at state 1 is clearly $V^*(1) = 2\gamma/(1-\gamma)$, while the actual return from the greedy policy π at 1 is $V^\pi(1) = 0$. Therefore

$$V^*(1) - V^\pi(1) = \frac{2\gamma}{1-\gamma} = \frac{2\gamma\varepsilon}{1-\gamma},$$

so the bound is attained.

The same Markov decision problem provides an example where the second bound is attained if we define

$$V(1) = V(2) = \frac{2}{1-\gamma}.$$

Since $V^*(1) = 2\gamma/(1-\gamma)$ and $V^*(2) = 2/(1-\gamma)$, we see that $V^*(1) < V(1)$ and $V^*(2) = V(2)$, so the condition $V^* \leq V$ is satisfied. In this case the V-residual at state 1 is

$$BV(1) - V(1) = \gamma V(1) - V(1) = \frac{2\gamma}{1-\gamma} - \frac{2}{1-\gamma} = -2,$$

while the V-residual at state 2 is

$$BV(2) - V(2) = 2 + \gamma V(2) - V(2) = 2 + \frac{2\gamma}{1-\gamma} - \frac{2}{1-\gamma} = 0,$$

so the V-error magnitude is $\varepsilon = 2$. The same policy π is greedy for this V, and we see that

$$V^*(1) - V^\pi(1) = \frac{2\gamma}{1-\gamma} = \frac{\gamma\varepsilon}{1-\gamma},$$

so the bound is attained in this case as well. □

The condition $V^* \leq V$ required for the second bound in this theorem may not always be easy to verify in practice, but the following result provides a sufficient condition for this that depends only on the V-residual.

Lemma 3.2 *Let V be a state value function. If $BV \leq V$, then $V^* \leq V$.*

Proof. This follows from two standard dynamic programming results whose easy proofs we omit. One is that B preserves the order relation on state value functions, which implies by induction that $B^n V \leq V$ for all n, and the other is that $V^* = \lim_{n\to\infty} B^n V$, for any state value function V, which is an easy consequence of Lemma 3.1 and the Bellman equation. □

We thus obtain the following more practical corollary to the second part of Theorem 3.2.

Corollary 3.1 *Let V be a value function on X, π a greedy policy for V, and $\varepsilon = \|BV - V\|$ the Bellman error magnitude for V. If $BV \leq V$ then*

$$V^\pi(x) \geq V^*(x) - \frac{\gamma\varepsilon}{1-\gamma}$$

for any state x. Furthermore, there is an example in which this bound is attained.

Proof. The bound follows immediately from Lemma 3.2 and Theorem 3.2. Furthermore, it is easily checked that the second example given in the proof of the Theorem 3.2 satisfies $BV \leq V$, so it serves as the claimed example here as well. $\square$

A useful way to interpret the above results is based on the observation that a constant immediate reward of r at every time step leads to an overall discounted reward of $r + \gamma r + \gamma^2 r + \ldots = r/(1-\gamma)$. Thus Theorem 3.2 says that a state value function V with V-error magnitude ε yields a greedy policy whose reward per step (on average) differs from optimal by at most $2\gamma\varepsilon$.

4 Results For State-Action Value Functions

In this section we give definitions and derive results analogous to those obtained above for the case when a state-action value function is used to determine a policy. Because state-action value functions are not as widely used in standard dynamic programming formulations as state value functions, we go into greater detail here than in the previous section.

4.1 Some Basic Definitions

A state-action value function Q is a function from $X \times A$ into the real numbers. For any stationary policy π, define Q^π to be that state-action value function assigning to state x and action a the quantity

$$Q^\pi(x,a) = E\left\{\sum_{t=0}^{\infty} \gamma^t r(x_t, a_t) \mid x_0 = x, a_0 = a\right\},$$

where it is also given that $x_{t+1} = f(x_t, a_t)$ and $a_t = \pi(x_t)$ for all $t > 0$. We further define Q^* to be Q^π for any optimal policy π. In addition, given Q and π, define $V_{Q,\pi}$ by $V_{Q,\pi}(x) = Q(x, \pi(x))$ and define V_Q by $V_Q(x) = \max_a Q(x,a)$.

We also define a partial order on state-action value functions by $Q \leq Q'$ if and only if $Q(x,a) \leq V'(x,a)$ for all $x \in X$ and $a \in A$.

4.2 Backup Operators

We define backup operators B^π and B on state-action value functions as follows:

$$\begin{aligned} B^\pi Q(x,a) &= E\left\{r(x,a) + \gamma V_{Q,\pi}(f(x,a))\right\} \\ &= R(x,a) + \gamma \sum_{y \in X} p_{xy}^a V_{Q,\pi}(y) \end{aligned}$$

and

$$\begin{aligned} BQ(x,a) &= E\left\{r(x,a) + \gamma V_Q(f(x,a))\right\} \\ &= R(x,a) + \gamma \sum_{y \in X} p_{xy}^a V_Q(y). \end{aligned}$$

While we use the same notation here as for the corresponding operators on state value functions, there will be no possibility of confusion since only the state-action value backup operators will be used in this section.

4.3 Greedy Policies

Given a state-action value function Q, define a stationary policy π to be *greedy* for Q if

$$\pi(x) = \arg\max_{a \in A} Q(x, a)$$

for any $x \in X$.

4.4 Maximum Norm Distance Measure

As with state value functions, we measure distances between state-action value functions according to the maximum norm, this time with

$$\|Q - Q'\| = \|Q - Q'\|_\infty = \max_{x \in X, a \in A} |Q(x, a) - Q'(x, a)|$$

for any state-action value functions Q and Q'.

4.5 Bellman Residual

Consider the equation $BQ = Q$. It has the same general form as the Bellman equation, and it is satisfied by $Q = Q^*$, as noted below in Lemma 4.2. Furthermore, it can also be shown that this solution is unique and that the equation $BQ^* = Q^*$ can be obtained as a direct consequence of the Bellman equation. Thus it might be appropriate to call $BQ = Q$ the *Bellman equation for state-action value functions.* Based on this reasoning, we define the *Bellman residual* for the state-action value function Q, or Q*-residual*, to be the state-action value function $BQ - Q$, and its maximum norm $\|BQ - Q\| = \max_{x,a} |BQ(x, a) - Q(x, a)|$ will be called the *Bellman error magnitude* for Q, or Q*-error magnitude.*

The significance of the Bellman residual, whether for a state value function or a state-action value function, was noted earlier. Here we make the additional observation that the individual components of the Q-residual are very closely related to what Q-learning tries to reduce toward zero. In particular, the TD errors used in the Q-learning algorithm are unbiased estimates of individual components of the Q-residual.

4.6 Some Preliminary Observations

Here we establish several elementary results relating the notions defined above. Since standard dynamic programming treatments generally do not deal with state-action value functions, we give detailed arguments justifying all results to be used later.

From the definitions of Q^π and V^π it is easy to see that

$$Q^\pi(x, a) = R(x, a) + \gamma \sum_{y \in X} p_{xy}^a V^\pi(y)$$

for any policy π. Furthermore, note that both Q^π and V^π are returns from similar policies (one of which is generally nonstationary) that differ only in the action applied at the outset. If the action applied at the outset at any given starting state x is selected according to policy π, then there is no difference. Thus, since $V_{Q^\pi,\pi}(x) = Q^\pi(x, \pi(x))$ for any x, it follows that $V_{Q^\pi,\pi} = V^\pi$. Specializing to the case when π is optimal, we also have

$$Q^*(x,a) = R(x,a) + \gamma \sum_{y \in X} p_{xy}^a V^*(y)$$

and $V_{Q^*} = V^*$.

But then, for any π, x, and a,

$$\begin{aligned} B^\pi Q^\pi(x,a) &= R(x,a) + \gamma \sum_{y \in X} p_{xy}^a V_{Q^\pi,\pi}(y) \\ &= R(x,a) + \gamma \sum_{y \in X} p_{xy}^a V^\pi(y) \\ &= Q^\pi(x,a). \end{aligned}$$

Therefore, $B^\pi Q^\pi = Q^\pi$. Also, for any x and a,

$$\begin{aligned} BQ^*(x,a) &= R(x,a) + \gamma \sum_{y \in X} p_{xy}^a V_{Q^*}(y) \\ &= R(x,a) + \gamma \sum_{y \in X} p_{xy}^a V^*(y) \\ &= Q^*(x,a), \end{aligned}$$

so $BQ^* = Q^*$.

We collect these observations in the following two lemmas.

Lemma 4.1 *For any state x and action a,*

$$Q^\pi(x,a) = R(x,a) + \gamma \sum_{y \in X} p_{xy}^a V^\pi(y)$$

and

$$Q^*(x,a) = R(x,a) + \gamma \sum_{y \in X} p_{xy}^a V^*(y).$$

□

Lemma 4.2 *Let π be any stationary policy. Then*

$$V_{Q^\pi,\pi} = V^\pi,$$

$$V_{Q^*} = V^*,$$

$$B^\pi Q^\pi = Q^\pi,$$

and

$$BQ^* = Q^*.$$

$\square$

For our next result, we need to make use of the following simple mathematical fact.

Lemma 4.3 *Let g_1 and g_2 be real-valued functions on a compact domain U. Then*

$$|\max_{u\in U} g_1(u) - \max_{u\in U} g_2(u)| \leq \max_{u\in U} |g_1(u) - g_2(u)|.$$

Proof. Pick $u_1 = \arg\max_{u\in U} g_1(u)$ and $u_2 = \arg\max_{u\in U} g_2(u)$. Consider first the case when $g_1(u_1) \geq g_2(u_2)$. Then

$$\begin{aligned}
|\max_{u\in U} g_1(u) - \max_{u\in U} g_2(u)| &= g_1(u_1) - g_2(u_2) \\
&\leq g_1(u_1) - g_2(u_1) \\
&= |g_1(u_1) - g_2(u_1)| \\
&\leq \max_{u\in U} |g_1(u) - g_2(u)|.
\end{aligned}$$

A symmetrical argument establishes the same result when $g_2(u_2) \geq g_1(u_1)$. $\square$

Lemma 4.4 *For any stationary policy π and any state-action value functions Q and Q',*

$$\|V_{Q,\pi} - V_{Q',\pi}\| \leq \|Q - Q'\|$$

and

$$\|V_Q - V_{Q'}\| \leq \|Q - Q'\|.$$

Proof. For any state x,

$$|V_{Q,\pi}(x) - V_{Q',\pi}(x)| = |Q(x,\pi(x)) - Q'(x,\pi(x))| \leq \|Q(x,\pi(x)) - Q'(x,\pi(x))\|.$$

Since this is true for all states, the first inequality follows.

For the second inequality, consider an arbitrary state x. Applying Lemma 4.3 yields

$$\begin{aligned}
|V_Q(x) - V_{Q'}(x)| &= |\max_a Q(x,a) - \max_a Q'(x,a)| \\
&\leq \max_a |Q(x,a) - Q'(x,a)| \\
&= \|Q - Q'\|,
\end{aligned}$$

and the desired result follows. $\square$

4.7 Derivation of Performance Bounds

Armed with the results developed above, we are now prepared to derive the desired tight performance bounds, which are given in Theorem 4.1 and Corollary 4.1. The progression of results we use corresponds to the derivation used for state value functions, beginning with the following contraction result.

Lemma 4.5 *For any stationary policy π and any state-action value functions Q and Q',*

$$\|B^\pi Q - B^\pi Q'\| \leq \gamma \|Q - Q'\|.$$

and

$$\|BQ - BQ'\| \leq \gamma \|Q - Q'\|$$

Proof. To prove the first inequality, we note that for any state x and action a,

$$\begin{aligned} B^\pi Q(x,a) - B^\pi Q'(x,a) &= R(x,a) + \gamma \sum_{y \in X} p_{xy}^a V_{Q,\pi}(y) - R(x,a) - \gamma \sum_{y \in X} p_{xy}^a V_{Q',\pi}(y) \\ &= \gamma \sum_{y \in X} p_{xy}^a \left[V_{Q,\pi}(y) - V_{Q',\pi}(y) \right], \end{aligned}$$

so

$$\begin{aligned} |B^\pi Q(x,a) - B^\pi Q'(x,a)| &\leq \gamma \sum_{y \in X} p_{xy}^a |V_{Q,\pi}(y) - V_{Q',\pi}(y)| \\ &\leq \gamma \sum_{y \in X} p_{xy}^a \|V_{Q,\pi} - V_{Q',\pi}\| \\ &= \gamma \|V_{Q,\pi} - V_{Q',\pi}\| \\ &\leq \gamma \|Q - Q'\|, \end{aligned}$$

where the last step is an application of the first part of Lemma 4.4.

The proof of the second inequality is obtained by following the same steps, but with B, V_Q, and $V_{Q'}$ replacing B^π, $V_{Q,\pi}$, and $V_{Q',\pi}$, respectively. In this case the last step is an application of the second part of Lemma 4.4. □

We also obtain the following result corresponding to Proposition 3.1.

Proposition 4.1 *For any state-action value functions Q and Q' and any policy π,*

$$\|Q - Q^\pi\| \leq \frac{\|Q - B^\pi Q\|}{1 - \gamma}$$

and

$$\|Q - Q^*\| \leq \frac{\|Q - BQ\|}{1 - \gamma}.$$

Proof. From the triangle inequality, the contraction property of B^π, and the fact that Q^π is fixed by B^π, we get

$$\|Q - Q^\pi\| \leq \|Q - B^\pi Q\| + \|B^\pi Q - Q^\pi\| \leq \|Q - B^\pi Q\| + \gamma \|Q - Q^\pi\|,$$

and the first inequality follows.

Similarly, the triangle inequality, the contraction property of B, and the fact that Q^* is fixed by B imply

$$\|Q - Q^*\| \leq \|Q - BQ\| + \|BQ - Q^*\| \leq \|Q - BQ\| + \gamma \|Q - Q^*\|,$$

and the second inequality follows. □

Theorem 4.1 *Let Q be a value function on $X \times A$ and let π be a greedy policy for Q. Let $\varepsilon = \|BQ - Q\|$ denote the Bellman error magnitude for Q. Then the actual return V^π from this policy satisfies*

$$V^\pi(x) \geq V^*(x) - \frac{2\varepsilon}{1-\gamma}$$

for any state x. Furthermore, if $V^* \leq V_Q$, then

$$V^\pi(x) \geq V^*(x) - \frac{\varepsilon}{1-\gamma}$$

for any state x. In addition, in each case there is an example where equality holds.

Proof. Since π is greedy for Q, $V_{Q^\pi} = V^\pi$. Also, $V_{Q^*} = V^*$. Together with the triangle inequality, this implies

$$\|V^* - V^\pi\| \leq \|V^* - V_Q\| + \|V_Q - V^\pi\| = \|V_{Q^*} - V_Q\| + \|V_Q - V_{Q^\pi}\|.$$

But then we can use Lemma 4.4 and Proposition 4.1 to conclude further that

$$\|V^* - V^\pi\| \leq \|Q^* - Q\| + \|Q - Q^\pi\| \leq \frac{\|Q - BQ\|}{1-\gamma} + \frac{\|Q - B^\pi Q\|}{1-\gamma}.$$

But when π is greedy for Q, $B^\pi Q = BQ$, so we get

$$\|V^* - V^\pi\| \leq \frac{2\|BQ - Q\|}{1-\gamma} = \frac{2\varepsilon}{1-\gamma}.$$

Finally, since $V^\pi \leq V^*$, this implies

$$V^*(x) - V^\pi(x) = |V^*(x) - V^\pi(x)| \leq \frac{2\varepsilon}{1-\gamma}$$

for any state x, from which the first bound follows.

To establish the second bound, note that $V^* \leq V_Q$ implies $V^\pi(x) \leq V^*(x) \leq V_Q(x)$ for any state x and for any policy π, so when π is greedy for Q,

$$\begin{aligned}
V^*(x) - V^\pi(x) &\leq V_Q(x) - V^\pi(x) \\
&= V_Q(x) - V_{Q^\pi}(x) \\
&\leq \|V_Q - V_{Q^\pi}\| \\
&\leq \|Q - Q^\pi\| \\
&\leq \frac{\|Q - B^\pi Q\|}{1-\gamma} \\
&\leq \frac{\|Q - BQ\|}{1-\gamma} \\
&= \frac{\varepsilon}{1-\gamma}.
\end{aligned}$$

To see that the first bound cannot be made tighter in general, consider a Markov decision problem having a single state 1 and two actions 1 and 2 which cause a self-transition at this state and which deterministically yield immediate rewards of 0 and 2, respectively. Clearly, $V^*(1) = 2/(1-\gamma)$. Now consider the state-action value function Q defined by

$$Q(1,1) = Q(1,2) = \frac{1}{1-\gamma}.$$

The stationary policy $\pi(1) = 1$ is a greedy policy for Q, and its return is clearly $V^\pi(1) = 0$, which differs from the optimal return by

$$V^*(1) - V^\pi(1) = \frac{2}{1-\gamma}.$$

Furthermore, $V_Q(1) = 1/(1-\gamma)$, so the Q-residual at $(1,1)$ is

$$BQ(1,1) - Q(1,1) = \gamma V_Q(1) - Q(1,1) = \frac{\gamma}{1-\gamma} - \frac{1}{1-\gamma} = -1.$$

A similar computation shows that the Q-residual at $(1,2)$ is 1, so the Q-error magnitude is $\varepsilon = 1$. Thus in this case, $V^*(1) - V^\pi(1)$ equals the upper bound $2\varepsilon/(1-\gamma)$.

The same Markov decision problem provides an example where the second bound is attained if we define

$$Q(1,1) = Q(1,2) = \frac{2}{1-\gamma}.$$

The condition $V^* \leq V_Q$ is satisfied since

$$V^*(1) = \frac{2}{1-\gamma} = \max_a Q(1,a) = V_Q(1).$$

The same policy $\pi(1) = 1$ is greedy for this Q, and yields the same difference $2/(1-\gamma)$ from the optimal return. In this case the Q-residual at $(1,1)$ is

$$BQ(1,1) - Q(1,1) = \frac{2\gamma}{1-\gamma} - \frac{2}{1-\gamma} = -2,$$

while the Q-residual at $(1, 1)$ is

$$BQ(1,2) - Q(1,2) = 2 + \frac{2\gamma}{1-\gamma} - \frac{2}{1-\gamma} = 0.$$

We see that the Q-error magnitude $\varepsilon = 2$, so $V^*(1) - V^\pi(1)$ equals the upper bound $\varepsilon/(1-\gamma)$. $\square$

A sufficient condition for $V^* \leq V_Q$ that may be easier to verify in practice is given by the following result.

Lemma 4.6 *Let Q be a state-action value function. If $BQ \leq Q$, then $V^* \leq V_Q$.*

Proof. Since $V^* = V_{Q^*}$, it is sufficient to show that $BQ \leq Q$ implies $Q^* \leq Q$ and that $Q' \leq Q$ implies $V_{Q'} \leq V_Q$ for any Q'. To prove the latter result, momentarily fix x and let $a^* = \arg\max_a Q'(x,a)$. Since $Q'(x,a) \leq Q(x,a)$ for any a,

$$\begin{aligned} V_{Q'}(x) &= \max_a Q'(x,a) \\ &= Q'(x,a^*) \\ &\leq Q(x,a^*) \\ &\leq \max_a Q(x,a) \\ &= V_Q(x). \end{aligned}$$

Since this holds for any x, $V_{Q'} \leq V_Q$.

Now we show $BQ \leq Q$ implies $Q^* \leq Q$, using an argument analogous to that used in the proof of Lemma 3.2. To do this, consider any state-action value function $Q' \leq Q$. By the result just obtained,

$$\begin{aligned} BQ'(x,a) &= R(x,a) + \gamma \sum_{y \in X} p^a_{xy} V_{Q'}(y) \\ &\leq R(x,a) + \gamma \sum_{y \in X} p^a_{xy} V_Q(y) \\ &= BQ(x,a) \end{aligned}$$

for any x and a, and we see that B preserves the order relation on state-action value functions. From this it follows by induction that $BQ \leq Q$ implies $B^n Q \leq Q$ for all n. In addition, because $BQ^* = Q^*$, we can use Lemma 4.5 inductively to conclude that $\|B^n Q - Q^*\| \leq \gamma^n \|Q - Q^*\|$ for any $n \geq 0$, which further implies $\lim_{n\to\infty} B^n Q = Q^*$. Thus $Q^* \leq Q$. $\square$

Combining this result with the second part of the theorem, we obtain the following result.

Corollary 4.1 *Let Q be a value function on $X \times A$, π a greedy policy for Q and $\varepsilon = \|BQ - Q\|$ the Bellman error magnitude for Q. Then $BQ \leq Q$ implies*

$$V^\pi(x) \geq V^*(x) - \frac{\varepsilon}{1-\gamma}$$

for any state x. Furthermore, there is an example in which this bound is attained.

Proof. The bound follows immediately from Lemma 4.6 and Theorem 4.1. Furthermore, it is easily checked that the second example given in the proof of Theorem 4.1 satisfies $BQ \leq Q$, so it serves as the claimed example here as well. □

As before, we can interpret these results in terms of reward per time step, with Theorem 4.1 saying that a state-action value function Q with Q-error magnitude ε yields a greedy policy whose reward per step (on average) differs from optimal by at most 2ε.

5 Discussion

A common practice in reinforcement learning applications is to work to minimize the Bellman residual (in the sense of trying to drive its components to zero) and then use the corresponding greedy policy. Most theoretical analyses of reinforcement learning have tended to rely, at least implicitly, on the idea that continued training leads eventually, if only in the limit, to solutions of the Bellman equation, and hence to optimal value functions. Singh and Yee (to appear) have taken the useful step of abandoning this idea, asking instead what would happen if one uses value functions that fail to satisfy the Bellman equation. However, their primary analysis has considered the case where this failure to satisfy the Bellman equation is measured in terms of distance from the unknown optimal value function, a more theoretical measure, rather than the Bellman equation error, which corresponds more directly with what most practical reinforcement learning algorithms actually try to minimize. They have also combined their primary bounds with the standard result given here as the second half of Proposition 3.1 to obtain bounds expressed in terms of this more practical measure. It is bounds of this type that should be emphasized since they provide a direct link between the Bellman equation error and the degree of nonoptimality of the resulting greedy policy and thus provide a very direct theoretical justification for the common practice cited above. The main difference between the Singh and Yee results of this type and those presented here is that the more indirectly derived bounds end up having an additional factor of $1-\gamma$ in the denominator, so the bounds derived here are significantly tighter.

The underlying motivation for performing the analyis presented here was to gain a better understanding of what to expect when a function approximator is used for the desired value function. The use of function approximators for this purpose is also common practice, and it is essential when the state space is large or continuous because it provides useful generalization based on a limited set of actually experienced transitions. But it is precisely in the case when a function approximator is used to represent a value function that it cannot generally be expected that it will be possible to drive the Bellman residual to zero at all states, so it is in this situation that the kind of results presented here are most meaningful. Nevertheless, the specific results given here fall short of addressing this situation fully, and we now consider what additional challenges remain to be faced and sketch some possible approaches for dealing with them.

Here we have assumed for simplicity that the state and action spaces are finite, but, interestingly, the main results, Theorems 3.2 and 4.1 and their corollaries, can be shown to carry over to continuous state spaces as well (with the maximum norm replaced by the supremum norm in general). However, in this case, or even when the state space is finite but large, the Bellman error magnitude ceases to be directly computable since it requires knowledge of the Bellman residual at all states. Thus, in this case, more theory is required to somehow relate the size of the Bellman error magnitude to the size of the Bellman residual at a reasonably small, finite subset of states, based on some assumptions on how the Bellman residual generalizes to nearby (or, more generally, "similar") states. The results given here further apply more generally to compact action spaces, but here again there is a gulf between what the theory deals with and what can be used in practice since it is not possible in general to compute the maximum over a continuous, or even finite but very large, action space. One possibility here is to develop an approach based not on truly greedy policies, that use the true maximum, but on ostensibly greedy policies that must make use of possibly flawed maximum-finding procedures.

While the complete theory necessary to recommend specific algorithms for these more general situations is lacking, certain novel possibilities are suggested by what is available so far. For example, the results given here express the performance bounds in terms of the L^∞ norm of the Bellman residual, while it is more typical to minimize its L^2 norm, as when backpropagation is used to train a feedforward neural net to approximate the desired value function. Of course, the bounds obtained here give rise to corresponding bounds expressed in terms of this L^2 norm, but it is interesting to consider what techniques may be available for minimizing the L^∞ norm directly. One such technique is to always identify the largest component of the Bellman residual and reduce it first, just as in the *priority-Dyna* variant (Moore & Atkeson, 1993; Peng, 1993; Peng & Williams, 1993) of Sutton's (1990; 1991) Dyna approach.[1] This interesting connection to an algorithm already studied and found useful in a somewhat different context deserves further study.

However, in the terminology of Singh (1993), such a technique relies on the use of *full backups*, which requires having the correct expected values over all stochastic possibilities occuring at a single step. It is not clear that there is any corresponding algorithm for minimizing the L^∞ norm of the Bellman residual that works on *sample backups*, of the kind that are used in Sutton's TD methods (1988).[2] Thus another useful extension of the results derived here is to the case when training is performed on a sample-by-sample basis in a stochastic environment. Such an extension is necessary to bring the theory into closer compliance with the way standard non-model-based TD methods are used in practice.

It should be noted that one limitation of standard reinforcement learning approaches that will not be helped by analyses like that presented here (or that makes such analyses impossible) is in situations when the optimal value function is not the only solution to the Bellman equation. Such a situation was encountered by Bradtke (1993) while considering the interplay of reinforcement learning methods with the use of non-tabular value function representations in the special case when the optimal value function can be represented exactly.

Finally, we note that the measure $V^*(x) - V^\pi(x)$ used here, as well as by Singh and Yee (to appear), to measure nonoptimality at state x, may not always be an appropriate measure for this,

---

[1] In general, when using a function approximator rather than a table-lookup representation, this reduction may have to be suitably small to insure that other components do not increase by too much, which means use of a small learning rate.

[2] In this regard it is interesting to note that the asymmetric aspect of these temporal difference methods, adjusting the left-hand side of the Bellman equation toward the current sample of the right-hand side but not vice-versa, appears to only be an issue because of the use of samples (cf. Werbos, 1990). When the actual Bellman equation is used, in which the expectation operator appears on the right-hand side, it is mathematically permissible to decrease the norm of the Bellman residual in a more symmetric fashion, although this may well give slower convergence to a minimum.

especially when discounting is used and nonzero rewards are sparse. This is demonstrated by the work of Thrun & Schwartz (1993) involving single-reward maze problems. In such problems, for any state far from the goal, the difference between the optimal discounted return and a nonoptimal return will be proportional to a high power of the discount factor and can thus be vanishingly small. This means that this measure can be small across all states even when there are many states for which the greedy policy fails to ever get to the goal.

References

Bertsekas, D. P. (1987). *Dynamic Programming: Deterministic and Stochastic Models.* Englewood Cliffs, NJ: Prentice Hall.

Bradtke, S. J. (1993). Reinforcement learing applied to linear quadratic regulation. In: S. J. Hanson, J. D. Cowan, & C. L. Giles (Eds.) *Advances in Neural Information Processing Systems 5.* San Mateo, CA: Morgan Kaufmann.

Moore, A. W. & Atkeson, C. G. (1993). Memory-based reinforcement learning: Converging with less data and less real time. In: S. J. Hanson, J. D. Cowan, & C. L. Giles (Eds.) *Advances in Neural Information Processing Systems 5.* San Mateo, CA: Morgan Kaufmann.

Peng, J. (1993). *Efficient Dynamic Programming-Based Learning for Control.* Ph.D. Dissertation, College of Computer Science, Northeastern University, Boston, MA.

Peng, J. & Williams, R. J. (1993). Efficient learning and planning within the Dyna framework, *Adaptive Behavior, 2*, 437-454.

Singh, S. P. (1993). *Learning Control in Dynamic Environments.* Ph.D. Dissertation, Department of Computer Science, University of Massachusetts, Amherst, MA.

Singh, S. P. & Yee, R. C. (To appear). An upper bound on the loss from approximate optimal-value functions. *Machine Learning.*

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning, 3*, 9-44.

Sutton, R. S. (1990). Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. *Proceedings of the Seventh International Conference in Machine Learning*, 216-224.

Sutton, R. S. (1991). Planning by incremental dynamic programming. *Proceedings of the 8th International Machine Learning Workshop.*

Thrun, S. & Schwartz, A. (1993) Issues in using function approximation for reinforcement learning. *Proceedings of the Fourth Connectionist Models Summer School.* Hillsdale, NJ: Erlbaum.

Watkins, C. J. C. H. (1989). *Learning from delayed rewards.* Ph.D. Dissertation, Cambridge University, Cambridge, England.

Watkins, C. J. C. H. & Dayan, P. (1992). Q-learning. *Machine Learning, 8*, 279-292.

Werbos, P. J. (1990). Consistency of HDP applied to a simple reinforcement learning problem. *Neural Networks, 3*, 179-189.