Tight Performance Bounds on Greedy Policies Based on Imperfect Value Functions

Ronald J. Williams
College of Computer Science, 161 CN
Northeastern University
Boston, MA 02115
rjw@ccs.neu.edu

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

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 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.

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).

For brevity, some of the straightforward mathematical proofs have been omitted in this paper. A longer version, containing all missing details, is also available (Williams & Baird, 1993). Also, Singh and Yee (to appear) have derived related bounds not as tight as those given here.

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 standard results from the theory of dynamic programming that provide a foundation for the arguments given here, 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 $\pi : X \to A$.

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

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

while BV assigns to state x the value

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

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, 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 for state value functions. The following lemma is an easy consequence of standard contraction results for discounted dynamic programming.

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

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

and

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

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*

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

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

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

while the second inequality of that lemma 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

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

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

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

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.

Now 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 $R(2,2) = 2$. Let the state value function V be 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, and it is straightforward to verify that the V-error magnitude is $\varepsilon = 1$ and

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

This shows that the first bound cannot be made tighter in general.

If, for the same Markov decision problem, we instead define

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

it is easy to see that $V^* \leq V$. The same policy π is greedy for this V, and it is straightforward to verify that the second bound is attained in this case. □

The condition $V^* \leq V$ required for the second bound in this theorem may not always be easy to

verify in practice, but it is easy to show that a sufficient condition for this to hold that depends only on the V-residual is that $BV \leq V$.

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 $r/(1-\gamma)$. Thus Theorem 3.1 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.

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:

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

and

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

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 easily shown that it is satisfied by $Q = Q^*$. 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 Derivation of Performance Bounds

Here we derive the desired tight performance bounds involving state-action value functions. First we state without proof two easily established lemmas that we will need in the proof of the main result. The first is an elementary result relating state-action value functions and certain state value functions derived from them, while the second corresponds to Lemma 3.1.

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

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

and

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

Lemma 4.2 *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

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

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

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

But then we can use Lemma 4.1 and Lemma 4.2 to conclude further that

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

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. Define the state-action value function Q by

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

The stationary policy $\pi(1) = 1$ is a greedy policy for Q, and it is straightforward to verify that the Q-error magnitude is $\varepsilon = 1$ and

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

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

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

It is easy to see that $V^* \leq V_Q$ in this case. The same policy π is greedy for this Q, and it is straightforward to verify that $V^*(1) - V^\pi(1)$ equals the upper bound $\varepsilon/(1-\gamma)$ in this case. □

Paralleling the state value case, a sufficient condition for $V^* \leq V_Q$ that may be easier to verify in practice is that $BQ \leq Q$.

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. There are several reasons it is not necessarily realistic to assume that this will lead to exact solutions to the Bellman equation. However, 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. By considering instead the kinds of approximate solutions one might expect to find through this typical reinforcement learning process, the analysis presented here provides a better theoretical justification for this practice.

One important reason this process may not lead to an exact solution of the Bellman equation is that a function approximator may be used for the desired value function. This is also common practice in reinforcement learning applications, 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.

Nevertheless, the specific results given here fall short of addressing this situation fully because of: (1) the impracticality of obtaining knowledge of the Bellman error magnitude at all states when the state space is large (or continuous); and (2) the impracticality of assuming that a greedy policy can be found when the action space is large (or continuous). The first difficulty suggests the need for more theory that relates 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. One recent approach to dealing with the second difficulty has been developed by Baird and Klopf (1993).

Finally, note that the error bounds derived here are worst-case bounds. The worst-case error of $2\varepsilon/(1-\gamma)$ is equivalent to adding an error of 2ε to the reinforcement on each time step. If ε is small relative to the typical one-step reinforcement, then the policy will be close to optimal. If the Bellman residual error is caused by errors in the function approximator for the value function, then one would expect that ε could be made fairly small relative to the size of a typical value. Unfortunately, for γ near one, the typical value can be much larger than the typical one-step reinforcement, so ε may be large relative to the one-step reinforcement. This suggests that in the worst case, when γ is near one, even small Bellman residual errors can accumulate and lead to highly suboptimal policies. The results might be more favorable and more generally applicable to the kinds of situations encountered in practice if the worst-case analysis performed here were replaced by a study of some suitably formulated notion of average-case behavior.

Acknowledgements

The first author was supported by Grant IRI-8921275 from the National Science Foundation and the second author was supported under Task 2312R102 by the Life and Environmental Sciences Directorate of the Air Force Office of Scientific Research.

References

Baird, L. C. & Klopf, A. H. (1993). *Reinforcement learning with high-dimensional, continuous actions.* U.S. Air Force Technical Report WL-TR-93-1147, Wright Laboratory, Wright-Patterson Air Force Base, OH.

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

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.

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.

Williams, R. J. & Baird, L. C., III (1993). *Tight performance bounds on greedy policies based on imperfect value functions* (Technical Report NU-CCS-93-14). Boston: Northeastern University, College of Computer Science.