An Upper Bound on the Loss from Approximate Optimal-Value Functions

Satinder P. Singh Richard C. Yee
Department of Computer Science
University of Massachusetts
Amherst, MA 01003
(413)-545-1596
singh@cs.umass.edu, yee@cs.umass.edu

Abstract

Many *reinforcement learning* (RL) approaches can be formulated from the theory of *Markov decision processes* and the associated method of *dynamic programming* (DP). The value of this theoretical understanding, however, is tempered by many practical concerns. One important question is whether DP-based approaches that use function approximation rather than lookup tables, can avoid catastrophic effects on performance. This note presents a result in Bertsekas (1987) which guarantees that small errors in the approximation of a task's *optimal value function* cannot produce arbitrarily bad performance when actions are selected greedily. We derive an upper bound on performance loss which is slightly tighter than that in Bertsekas (1987), and we show the extension of the bound to *Q-learning* (Watkins, 1989). These results provide a theoretical justification for a practice that is common in reinforcement learning.

Keywords: Reinforcement Learning, Approximation, Theoretical Analysis
Running Head: Loss from Approximate Optimal-Value Functions

1 Introduction

Recent progress in *reinforcement learning* has been made by forming connections to the theory of *Markov decision processes* and the associated optimization method of *dynamic programming* (DP) (Barto et al., 1990; Barto et al., 1991; Sutton, 1988; Watkins, 1989; Sutton, 1990). Theoretical results guarantee that many DP-based learning methods will find optimal solutions for a wide variety of search, planning, and control problems. Unfortunately, such results often do not sufficiently constrain the computational resources required. In particular, DP-based methods form *value functions* which assign numeric estimates of utility to task states. A common theoretical assumption is that such functions are implemented as lookup tables, i.e., that all elements of the function's domain are individually represented and updated (e.g., Sutton, 1988; Watkins & Dayan, 1992; Barto et al., 1991; however, see Bertsekas, 1987, and Bradtke, Forthcoming, for approximation results in restricted domains). If practical concerns dictate that value functions be approximated, how might performance be effected? Is it possible that, despite some empirical evidence to the contrary (e.g., Barto et al., 1983; Anderson, 1986; Tesauro, 1992), small errors in approximations could result in arbitrarily bad performance in principle? If so, this could raise significant concerns about the use of function approximation in DP-based learning.

This note presents to the machine learning community a result in Bertsekas (1987)[1] which guarantees that a good approximation of a task's *optimal value function* will yield reasonable performance when actions are selected greedily. Using a natural definition of the *loss* in performance due to approximation, we derive an upper bound on the loss which is slightly tighter than the one indicated in Bertsekas (1987). We also show the corresponding extension to *Q-learning* (Watkins, 1989). Note that these results do not address the issue of *converging* to good approximations. They show that if good approximations of values are achieved, then reasonable performance can be guaranteed.

[1]We thank Andrew Barto for identifying that our result for Q-value functions was a simple extension to Bertsekas's (1987) result on the usual value functions in dynamic programming theory.

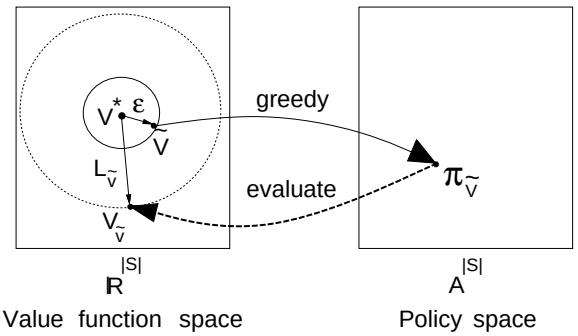

Figure 1: Given $\tilde{V}$, an approximation within $\epsilon > 0$ of V^*, derive the corresponding greedy policy $\pi_{\tilde{V}}$. The resulting loss in value, $V^* - V_{\tilde{V}}$, is bounded above by $(2\gamma\epsilon)/(1-\gamma)$.

2 Problem Statement and Theorem

We consider stationary Markovian decision tasks (MDTs) that have finite state and action sets (e.g., see Bertsekas, 1987, Barto et al., 1990). Let S be the state set, $A(x)$ be the action set for state $x \in S$, and $P_{xy}(a)$ be the probability of transiting from state x to state y on the execution of action $a \in A(x)$. Furthermore, we consider only infinite-horizon tasks with geometric discounting of payoffs. Let $R(x,a)$ be the expected payoff received on executing action a in state x, and let $\gamma \in [0,1)$ be the discount factor. The performance of a stationary control policy $\pi : S \rightarrow A$ is measured by a value function $V_\pi : S \rightarrow \Re$, defined by

$$\begin{aligned} V_\pi(x) &\stackrel{def}{=} E_\pi \left\{ \sum_{t=0}^{\infty} \gamma^t r_t \mid x_0 = x \right\} \\ &= R\left(x,\, \pi(x)\right) + \gamma \sum_{y \in S} P_{xy}\left(\pi(x)\right)\, V_\pi(y), \end{aligned}$$

where x_t and r_t denote respectively the state and payoff received at time t, and E_π is the expectation operator given that all actions are selected according to policy π. An *optimal policy* yields values that are greater than or equal to those of any other policy, i.e., if π^* is an optimal policy and V^* is its value function, then for all policies $\pi, \forall x,\, V^*(x) \geq V_\pi(x)$. V^* is the *optimal value function*, and it is unique for MDTs of the indicated type.

Policies can also be derived from value functions in a straightforward way.

Given value function $\tilde{V}$, a *greedy policy* $\pi_{\tilde{V}}$ can be defined as

$$\pi_{\tilde{V}}(x) \stackrel{def}{=} \arg\max_{a \in A(x)} \left[R(x,a) + \gamma \sum_{y \in S} P_{xy}(a)\, \tilde{V}(y) \right],$$

where ties for the maximum action are broken arbitrarily. Figure 1 illustrates the relationship between the evaluation of policies and the derivation of greedy policies. A greed policy $\pi_{\tilde{V}}$ gives rise to its own value function $V_{\pi_{\tilde{V}}}$, denoted simply as $V_{\tilde{V}}$. In general, $\tilde{V} \neq V_{\tilde{V}}$, i.e., the value function used for deriving a greedy policy is different than the value function resulting from the policy's evaluation. Equality between $\tilde{V}$ and $V_{\tilde{V}}$ occurs if and only if $\tilde{V}$ is optimal, in which case any greedy policy will be optimal.

For a greedy policy $\pi_{\tilde{V}}$ derived from $\tilde{V}$, an approximation to V^*, define the *loss function* $L_{\tilde{V}}$ such that $\forall x \in S$,

$$L_{\tilde{V}}(x) \stackrel{def}{=} V^*(x) - V_{\tilde{V}}(x),$$

$L_{\tilde{V}}(x)$ is the expected loss in the value of state x resulting from the use of policy $\pi_{\tilde{V}}$ instead of an optimal policy. The following theorem gives an upper bound on the loss $L_{\tilde{V}}$.

Theorem 1: (Cf. Bertsekas, 1987, p. 236, #14(c): the following bound is tighter by a factor of γ.) Let V^* be the optimal value function for a discrete-time MDT having finite states and actions and an infinite, geometrically discounted horizon, $\gamma \in [0,1)$. If $\tilde{V}$ is a function such that $\forall x \in S$, $\left| V^*(x) - \tilde{V}(x) \right| \leq \epsilon$, and $\pi_{\tilde{V}}$ is a greedy policy for $\tilde{V}$, then $\forall x$,

$$L_{\tilde{V}}(x) \leq \frac{2\gamma\epsilon}{1-\gamma}.$$

One interpretation of this result is that if the approximation to the optimal value function is off by no more than ϵ, then the average worst-case loss per time step cannot be more than $2\gamma\epsilon$, under a greedy policy.

proof: There exists a state z that achieves the maximum loss: $\exists z \in S$, $\forall x \in S$, $L_{\tilde{V}}(z) \geq L_{\tilde{V}}(x)$. For state z consider an optimal action, $a = \pi^*(z)$, and

the action of $\pi_{\tilde{V}}$, $b = \pi_{\tilde{V}}(z)$. Because $\pi_{\tilde{V}}$ is a greedy policy for $\tilde{V}$, b must appear at least as good as a:

$$\begin{aligned} R(z,a) + \gamma \sum_{y \in S} P_{zy}(a)\, \tilde{V}(y) &\leq R(z,b) + \gamma \sum_{y \in S} P_{zy}(b)\, \tilde{V}(y) \quad (1) \\ R(z,a) - R(z,b) &\leq \gamma \sum_{y} \left[P_{zy}(b)\ (V^*(y) + \epsilon) - P_{zy}(a)\ (V^*(y) - \epsilon) \right] \\ R(z,a) - R(z,b) &\leq 2\gamma\epsilon + \gamma \sum_{y} \left[P_{zy}(b)\, V^*(y) - P_{zy}(a)\, V^*(y) \right]. \quad (2) \end{aligned}$$

The maximal loss is

$$\begin{aligned} L_{\tilde{V}}(z) &= V^*(z) - V_{\tilde{V}}(z) \\ &= R(z,a) - R(z,b) + \gamma \sum_{y} \left[P_{zy}(a) V^*(y) - P_{zy}(b) V_{\tilde{V}}(y) \right]. \quad (3) \end{aligned}$$

Substituting from (2) gives

$$\begin{aligned} L_{\tilde{V}}(z) &\leq 2\gamma\epsilon + \gamma \sum_{y} \left[P_{zy}(b)\, V^*(y) - P_{zy}(a)\, V^*(y) \right. \\ &\qquad \left. + P_{zy}(a)\, V^*(y) - P_{zy}(b)\, V_{\tilde{V}}(y) \right] \\ L_{\tilde{V}}(z) &\leq 2\gamma\epsilon + \gamma \sum_{y} P_{zy}(b)\ \left[V^*(y) - V_{\tilde{V}}(y) \right] \\ L_{\tilde{V}}(z) &\leq 2\gamma\epsilon + \gamma \sum_{y} P_{zy}(b)\, L_{\tilde{V}}(y) \end{aligned}$$

Because, by assumption, $\forall y \in S$, $L_{\tilde{V}}(z) \geq L_{\tilde{V}}(y)$,

$$\begin{aligned} L_{\tilde{V}}(z) &\leq 2\gamma\epsilon + \gamma \sum_{y} P_{zy}(b)\, L_{\tilde{V}}(z) \\ L_{\tilde{V}}(z) &\leq \frac{2\gamma\epsilon}{1-\gamma} \end{aligned}$$

Q.E.D.

This result extends to a number of related cases.

Approximate payoffs Theorem 1 assumes that the expected payoffs are known exactly. If the true expected payoffs $R(x,a)$ are approximated by $\tilde{R}(x,a)$, the upper bound on the loss is as follows.

Corollary 1: If $\forall x \in S$, $\left|V^*(x) - \tilde{V}(x)\right| \leq \epsilon$ and $\forall a \in A(x)$, $\left|R(x,a) - \tilde{R}(x,a)\right| \leq \alpha$, then $\forall x$,

$$L_{\tilde{V}}(x) \leq \frac{2\gamma\epsilon + 2\alpha}{1-\gamma},$$

where $\pi_{\tilde{V}}$ is the greedy policy for $\tilde{V}$.

proof: Inequality (1) becomes

$$\tilde{R}(z,a) + \gamma \sum_{y \in S} P_{zy}(a)\, \tilde{V}(y) \leq \tilde{R}(z,b) + \gamma \sum_{y \in S} P_{zy}(b)\, \tilde{V}(y),$$

and (2) becomes

$$R(z,a) - R(z,b) \leq 2\gamma\epsilon + 2\alpha + \gamma \sum_{y} \left[P_{zy}(b)\, V^*(y) - P_{zy}(a)\, V^*(y)\right].$$

Substitution into (3) yields the bound. Q.E.D.

Q-learning If neither the payoffs nor the state-transition probabilities are known, then the analogous bound for *Q-learning* (1989) is as follows. Evaluations are defined by

$$Q_\pi(x_t, a) \stackrel{def}{=} R(x_t, a) + \gamma E\left\{V_\pi(x_{t+1})\right\},$$

where $V_\pi(x) = \max_a Q_\pi(x,a)$. Given function $\tilde{Q}$, the greedy policy $\pi_{\tilde{Q}}$ is defined by

$$\pi_{\tilde{Q}}(x) \stackrel{def}{=} \arg\max_{a \in A(x)} \tilde{Q}(x,a).$$

The loss is then expressed as

$$L_{\tilde{Q}}(x) \stackrel{def}{=} Q^*(x,\, \pi^*(x)) - \tilde{Q}(x,\, \pi_{\tilde{Q}}(x)).$$

Corollary 2: If $\forall x \in S$, $\forall a \in A(x)$, $\left|Q^*(x,a) - \tilde{Q}(x,a)\right| \leq \epsilon$, then $\forall x$,

$$L_{\tilde{Q}}(x) \leq \frac{2\epsilon}{1-\gamma}.$$

proof: Inequality (1) becomes $\tilde{Q}(z,a) \leq \tilde{Q}(z,b)$, which gives

$$\begin{aligned} Q^*(z,a) - \epsilon &\leq Q^*(z,b) + \epsilon \\ R(z,a) + \gamma \sum_y P_{zy}(a)\, V^*(y) - \epsilon &\leq R(z,b) + \gamma \sum_y P_{zy}(b)\, V^*(y) + \epsilon \\ R(z,a) - R(z,b) &\leq 2\epsilon + \gamma \sum_y \left[P_{zy}(b)\, V^*(y) - P_{zy}(a)\, V^*(y) \right]. \end{aligned}$$

Substitution into (3) yields the bound. Q.E.D.

3 Discussion

The bounds of the theorem and its corollaries guarantee that the performance of DP-based learning approaches will not be far from optimal if (a) good approximations to optimal value functions are achieved, (b) a corresponding greedy policy is followed, and (c) the discount factor, γ, is not too close to 1.0. The analogous result holds for *indefinite* horizon, undiscounted MDTs (as defined in Barto *et al.* 1991). Although this result does not address convergence, it nevertheless helps to validate many practical approaches to DP-based learning that use approximations.

The theorem also does not directly address policy-derivation methods other than the indicated greedy one. For other methods, it is not clear, in general, what criteria to place on approximations of value functions because the criteria may depend upon the specifics of a derivation method. Greedy policy derivation allows one to specify an error-criterion on approximations, i.e., that they be within ϵ of optimal, under the max norm.

Acknowledgement

We thank Andrew Barto for identifying the connection to (Bertsekas, 1987). This work was supported by a grant (ECS-9214866) to Prof. A. G. Barto from the National Science Foundation.

References

Anderson, C. W. (1986). *Learning and Problem Solving with Multilayer Connectionist Systems.* PhD thesis, University of Massachusetts, Department of Computer and Information Science, University of Massachusetts, Amherst, MA 01003.

Barto, A. G., Bradtke, S. J., & Singh, S. P. (1991). Real-time learning and control using asynchronous dynamic programming. Technical Report TR-91-57, Department of Computer Science, University of Massachusetts.

Barto, A. G., Sutton, R. S., & Anderson, C. W. (1983). Neuronlike elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, *13*(5), 834–846.

Barto, A. G., Sutton, R. S., & Watkins, C. J. C. H. (1990). Learning and sequential decision making. In Gabriel, M. & Moore, J. (Eds.), *Learning and Computational Neuroscience: Foundations of Adaptive Networks*, chapter 13, pages 539–602. Cambridge, MA: Bradford Books/MIT Press.

Bertsekas, D. P. (1987). *Dynamic programming: Deterministic and stochastic models.* Englewood Cliffs, NJ: Prentice Hall.

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. In Porter, B. W. & Mooney, R. H. (Eds.), *Machine Learning: Proceedings of the Seventh International Conference (ML90)*, pages 216–224, San Mateo, CA. Morgan Kaufmann.

Tesauro, G. (1992). Practical issues in temporal difference learning. *Machine Learning*, *8*(3/4), 257–277.

Watkins, C. J. C. H. (1989). *Learning from Delayed Rewards.* PhD thesis, King's College, University of Cambridge, Cambridge, England.

Watkins, C. J. C. H. & Dayan, P. (1992). Q-learning. *Machine Learning*, *8*(3/4), 279–292.