Machine Learning, 16, 227–233 (1994)

Technical Note

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

SATINDER P. SINGH* singh@cs.umass.edu
RICHARD C. YEE yee@cs.umass.edu
Department of Computer Science, University of Massachusetts, Amherst, MA 01003

Editor: Richard Sutton

Abstract. Many reinforcement learning approaches can be formulated using 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 of 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 by a greedy policy. We derive an upper bound on performance loss that 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 partial theoretical rationale for the approximation of value functions, an issue of great practical importance in reinforcement learning.

Keywords: Reinforcement learning, Markov decision processes, function approximation, performance loss

1. Introduction

Recent progress in reinforcement learning has been made by forming connections to the theory of *Markov decision processes* (MDPs) and the associated optimization method of *dynamic programming* (DP) (Barto et al., 1990; Barto et al., 1991; Sutton, 1988; Watkins, 1989; Sutton, 1990; Werbos, 1987). 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 fail to assume practical limitations on 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, 1993, for approximation results for restricted classes of MDPs). If practical concerns dictate that value functions must be approximated, how might performance be affected? 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? If so, this could raise significant concerns about the use of function approximation in DP-based learning.

[*] Singh's address from September 1993 to August 1995 will be: Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, e-mail: singh@psyche.mit.edu.

This note presents to the machine learning community a result of Bertsekas (1987) which guarantees that a good approximation of a task's optimal value function will yield reasonable performance when actions are selected according to a greedy policy. 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). Although 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.

2. Problem statement and theorem

We consider stationary Markovian decision processes (MDPs, henceforth also called *tasks*) that have finite state and action sets (e.g., see Bertsekas, 1987; Barto et al., 1990). Let X be the state set, $A(x)$ be the action set for state $x \in X$, and $P_{xy}(a)$ be the probability of a transition from state x to state y, given the execution of action $a \in A(x)$. Let $R(x, a)$ be the expected payoff received on executing action a in state x. We consider only stationary deterministic policies, $\pi: X \rightarrow A$, and infinite-horizon tasks with geometrically discounted payoffs, $\gamma \in [0, 1)$. A value function is any real-valued function of states, $V: X \rightarrow \Re$. In particular, value function V_π measures policy π's performance if, for all $x \in X$,

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

where x_t and r_t respectively denote the state and payoff received at time t, and E_π is the expectation given that actions are selected according to policy π. The determination of V_π for a given π is called *policy evaluation.*

The value function for an *optimal policy* is greater than or equal to that of any other policy, i.e., if π^* is an optimal policy and V^* is its value function, then for all policies $\pi, V^*(x) \geq V_\pi(x)$, for all $x \in X$. V^* is the *optimal value function*, and it is unique for this class of MDPs.

Value functions can also give rise to policies in a straightforward fashion. Given value function $\tilde{V}$, a *greedy policy* $\pi_{\tilde{V}}$ can be defined by selecting for each state the action that maximizes the state's value, i.e.,

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

where ties for the maximum action are broken arbitrarily. Evaluating a greedy policy $\pi_{\tilde{V}}$ yields a new value function $V_{\pi_{\tilde{V}}}$, which we abbreviate as $V_{\tilde{V}}$. Figure 1 illustrates the relationship between the derivation of greedy policies and policy evaluation. Value function $\tilde{V}$ gives rise to greedy policy $\pi_{\tilde{V}}$ which, when evaluated, yields $V_{\tilde{V}}$. In general, $\tilde{V} \neq V_{\tilde{V}}$. Equality occurs if and only if $\tilde{V} = V^*$, in which case any greedy policy will be optimal.

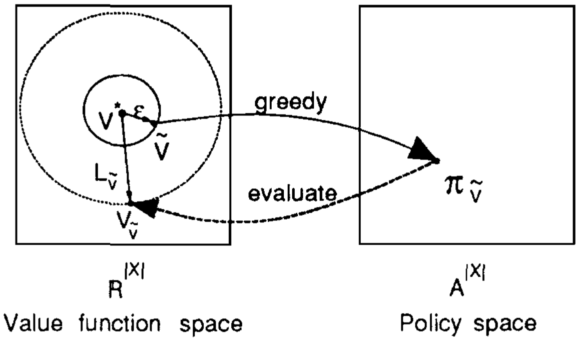

Figure 1. Loss from approximate optimal-value functions. Given $\tilde{V}$, an approximation within $\epsilon > 0$ of V^*, derive the corresponding greedy policy $\pi_{\tilde{V}}$. The resulting loss in value, $V^* - V_{\tilde{V}}$, is bounded above by $(2\gamma\epsilon)/(1-\gamma)$.

For a greedy policy $\pi_{\tilde{V}}$ derived from $\tilde{V}$ define the *loss function* $L_{\tilde{V}}$ such that for all $x \in X$,

$$L_{\tilde{V}}(x) \overset{\text{def}}{=} V^*(x) - V_{\tilde{V}}(x).$$

$L_{\tilde{V}}(x)$ 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. *Let V^* be the optimal value function for a discrete-time MDP having finite state and action sets and an infinite horizon with geometric discounting: $\gamma \in [0,1)$. If $\tilde{V}$ is a function such that for all $x \in X$, $|V^*(x) - \tilde{V}(x)| \leq \epsilon$, and $\pi_{\tilde{V}}$ is a greedy policy for $\tilde{V}$, then for all x,*

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

(*Cf. Bertsekas,* 1987, *p.* 236, #14(*c*)*: the preceding bound is tighter by a factor of* γ.)

Proof: There exists a state that achieves the maximum loss. Call this state z. Then for all $x \in X$, $L_{\tilde{V}}(z) \geq L_{\tilde{V}}(x)$. For state z consider an optimal action, $a = \pi^*(z)$, and the action specified by $\pi_{\tilde{V}}$, $b = \pi_{\tilde{V}}(z)$. Because $\pi_{\tilde{V}}$ is a greedy policy for $\tilde{V}$, b must appear at least as good as a:

$$R(z,a) + \gamma \sum_{y \in X} P_{zy}(a)\tilde{V}(y) \leq R(z,b) + \gamma \sum_{y \in X} P_{zy}(b)\tilde{V}(y). \qquad (1)$$

Because for all $y \in X$, $V^*(y) - \epsilon \leq \tilde{V}(y) \leq V^*(y) + \epsilon$,

$$R(z,a) + \gamma \sum_{y \in X} P_{zy}(a)(V^*(y) - \epsilon) \leq R(z,b) + \gamma \sum_{y \in X} P_{zy}(b)(V^*(y) + \epsilon).$$

Therefore, we have that

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

The maximal loss is

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

Substituting from (2) gives

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

Because, by assumption, $L_{\tilde{V}}(z) \geq L_{\tilde{V}}(y)$, for all $y \in X$, we have

$$L_{\tilde{V}}(z) \leq 2\gamma\epsilon + \gamma \sum_y P_{zy}(b)L_{\tilde{V}}(z).$$

Simplifying yields

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

This result extends to a number of related cases.

Approximate payoffs. The theorem assumes that the expected payoffs are known exactly. If the true expected payoff $R(x,a)$ is approximated by $\tilde{R}(x,a)$, for all $x \in X$ and $a \in A(x)$, then the upper bound on the loss is as follows.

Corollary 1. *If for all* $|V^*(x) - \tilde{V}(x)| \leq \epsilon$, *for all* $x \in X$, *and* $|R(x,a) - \tilde{R}(x,a)| \leq \alpha$, *for all* $a \in A(x)$, *then*

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

for all $x \in X$, *where* $\pi_{\tilde{V}}$ *is the greedy policy for* $\tilde{V}$.

Proof: Inequality (1) becomes

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

and (2) becomes

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

Substitution into (3) yields the bound. ■

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

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

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

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

The loss is then expressed as

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

COROLLARY 2. *If* $|Q^*(x,a) - \tilde{Q}(x,a)| \le \epsilon$, *for all* $x \in X$ *and* $a \in A(x)$, *then for all* $x \in X$,

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

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

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

Substitution into (3) yields the bound. ■

Bounding ϵ. As Williams and Baird have pointed out, the bounds of the preceding theorem and corollaries cannot be computed in practice because the determination of ϵ requires knowledge of the optimal value function, V^*. Nevertheless, upper bounds on approximation losses can be computed from the following upper bound on ϵ (Porteus, 1971). Let

$$\begin{aligned} C(x) &\stackrel{\text{def}}{=} V'(x) - \tilde{V}(x) \\ &= \max_{a \in A(x)} \left[R(x,a) + \sum_{y \in X} P_{xy}(a)\tilde{V}(y) \right] - \tilde{V}(x), \end{aligned} \tag{4}$$

and let $\delta = \max_{x \in X} C(x)$; then $\epsilon \leq \frac{\delta}{1-\gamma}$. Replacing ϵ by $\frac{\delta}{1-\gamma}$ in the bounds of the theorem and corollaries yields new bounds expressed in terms of a quantity, δ, that can be computed from successive value function approximations, V' and $\tilde{V}$ of Equation (4), which arise naturally in DP algorithms such as value iteration. In model-free algorithms such as Q-learning, δ can be stochastically approximated. See Williams and Baird (1993) for the derivation of tighter bounds of this type.

3. Discussion

The theorem and its corollaries guarantee that the infinite-horizon sum of discounted payoffs accumulated by 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. More specifically, the bounds can be interpreted as showing that greedy policies based on approximations can do no worse than policies whose expected loss at each time step is about twice the approximation error.

It should be pointed out that incurring only "small losses" in the sum of discounted payoffs need not always correspond to the intuitive notion of "near success" in a task. For example, if a task's sole objective is to reach a goal state, then a sufficiently small discount factor might yield only a small difference between a state's value under a policy that would lead optimally to the goal and the state's value under a policy that would *never* lead to the goal. In such cases, care must be taken in formulating tasks, e.g., in choosing the magnitudes of payoffs and discount factors. One must try to ensure that policies meeting important performance criteria will be learned robustly in the face of small numerical errors.

Although the above bounds on the loss function can help to justify DP-based learning approaches that do not implement value functions as lookup tables, there are currently few theoretical guarantees that such approaches will, in fact, obtain good approximations to optimal-value functions (i.e., small values of ϵ, or δ). Indeed, informal reports by researchers indicate that it can be quite difficult to achieve success with DP-based approaches that incorporate common function approximation methods. Thus, the theoretical and empirical investigation of function approximation and DP-based learning remains an active area of research.

Acknowledgments

We thank Andrew Barto for identifying the connection to (Bertsekas, 1987), and we thank the anonymous reviewers for many helpful comments. This work was supported by grants to Prof. A.G. Barto from the Air Force Office of Scientific Research, Bolling AFB, under Grant AFOSR-F49620-93-1-0269 and by the National Science Foundation under Grants ECS-8912623 and ECS-9214866.

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., and 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., and 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., and Watkins, C.J.C.H. (1990). Learning and sequential decision making. In M. Gabriel and J. Moore (Eds.), *Learning and Computational Neuroscience: Foundations of Adaptive Networks*, chapter 13. Cambridge, MA: Bradford Books/MIT Press.

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

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

Porteus, E. (1971). Some bounds for discounted sequential decision processes. *Management Science, 19*, 7–11.

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 B.W. Porter and R.H. Mooney (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. and Dayan, P. (1992). Q-learning. *Machine Learning, 8*(3/4), 279–292.

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

Werbos, P.J. (1987). Building and understanding adaptive systems: A statistical/numerical approach to factory automation and brain research. *IEEE Transactions on Systems, Man, and Cybernetics, 17*(1), 7–20.

Williams, R.J. and Baird, L.C. (1993). Analysis of some incremental variants of policy iteration: First steps toward understanding actor-critic learning systems. Technical Report NU-CCS-93-11, Northeastern University, College of Computer Science, Boston, MA 02115.

Received May 6, 1993
Final Manuscript January 5, 1994