The Loss from Imperfect Value Functions in Expectation-Based and Minimax-Based Tasks

MATTHIAS HEGER

heger@informatik.uni-bremen.de

Zentrum für Kognitionswissenschaften, Universität Bremen, FB3 Informatik, Postfach 330 440, 28334 Bremen, Germany

Editor: Leslie Pack Kaelbling

Abstract. Many reinforcement learning (RL) algorithms approximate an optimal value function. Once the function is known, it is easy to determine an optimal policy. For most real-world applications, however, the value function is too complex to be represented by lookup tables, making it necessary to use function approximators such as neural networks. In this case, convergence to the optimal value function is no longer guaranteed and it becomes important to know to which extent performance diminishes when one uses approximate value functions instead of optimal ones. This problem has recently been discussed in the context of expectation-based Markov decision problems. Our analysis generalizes this work to minimax-based Markov decision problems, yields new results for expectation-based tasks, and shows how minimax-based and expectation-based Markov decision problems relate.

Keywords: Reinforcement Learning, Dynamic Programming, Performance Bounds, Minimax Algorithms, Q-Learning

1. Introduction

Reinforcement learning (RL) is learning to solve decision problems from experience: an agent interacts with its environment (world) and receives reinforcement signals as punishments for its actions. Its task is to find a behavior that minimizes the punishment, specified as rules that tell the agent which action to choose in every possible situation. (The reinforcement signals can also be interpreted as rewards. This is equivalent to their interpretation as punishments, except that the agent now has to maximize the reward.)

The interaction of the agent with its environment is often modeled as a special kind of stochastic process, namely a Markov decision process (MDP). S is the set of states of the environment as the agent perceives it, A the set of actions that the agent has available, and C the set of scalar reinforcement signals, which we call immediate costs because they can be interpreted as the effort required to execute the actions. In this paper, we limit ourselves to finite sets S and A that are subsets of $\mathcal{N}$. We assume that the elements of $C \subset \mathcal{R}$ are countable, bounded, and nonnegative. Furthermore, the agent does not have all actions available in every state. The nonempty sets $A(i) \subseteq A$ denote the set of admissible actions in states i .

The interaction of the agent with its environment takes place in episodes, where episodes correspond to time steps: First, the agent observes the *starting state* $i \in S$. It then has to execute an *action* $a \in A(i)$, which causes a state transition from state i to a *successor state* $j \in S$. Finally, the agent receives the *reinforcement signal* $r \in C$.

MDPs make the important assumption that the probability distribution over the successor states depends only on the starting state i and the executed action a but not on t or any previous episodes. $P_S(i, a, j)$ denotes the probability that the successor state of an episode is j if action a is executed in starting state s . Similarly, MDPs assume that the probability that the immediate cost of episode t equals a given number $r \in C$ depends only on the starting state, the executed action, and the successor state of that episode, but neither on t nor past episodes. $P_C(i, a, j, r)$ denotes the probability that the immediate cost of an episode is r for a given starting state i , action a , and successor state j . Furthermore, MDPs assume that the starting state of an episode is identical to the successor state of the previous episode. If the MDP is deterministic, then the successor state j_i^a and the immediate cost c_i^a are uniquely determined by the starting state i and the executed action a .

The behavior of the agent is specified by a *policy*, which is a mapping from situations to actions. In general, the action could depend on the current time, the current state of the agent, and all previous states and actions. Furthermore, the action could be selected probabilistically. *Stationary policies* determine which action to execute based only on the current state of the agent, but prove to be very powerful.

The states and immediate costs are random variables because of the probabilistic nature of the immediate costs and state transitions. These random variables essentially depend on the policy that the agent follows. Therefore, we use the following notation: I_t^π denotes the starting state and C_t^π the immediate cost, where π and t describe the policy of the agent and the time index of the episode, respectively.

There are two problems that make it difficult to define what an optimal policy is. The first problem is often called *delayed reinforcement*, because actions may reveal their consequences only many time steps after they have been executed. Hence, the action that has the lowest immediate cost is not necessarily best from a global point of view. Therefore, one should consider all immediate costs, including the future costs, which is usually done by evaluating policies according to their return

$$R_\gamma^\pi = \sum_{\tau=0}^{\infty} \gamma^\tau C_\tau^\pi$$

where $0 \leq \gamma < 1$ is a discount factor that keeps the sum finite and has the effect that costs are weighted less if they are obtained farther into the future.

In deterministic domains, the return of a policy can already be used to measure its performance. In this case, a policy is optimal if it minimizes the return. Unfortunately, there is a second problem, since the return is usually not a real number, but a random variable. In probabilistic domains, using the same policy repeatedly in the same starting state can result in different returns. The common way to handle this problem is to use the expected value of the return as performance measure (Barto, Sutton & Anderson, 1983; Watkins, 1989). According to this so-called expected value criterion, a policy is optimal if it minimizes the expected value of the return. In operations research and decision theory, however, it is well known that it is not always reliable and can even be very misleading to use the expected value as decision criterion (e.g., Taha, 1987). In (Heger, 1994a), Heger presented dynamic programming algorithms and a RL algorithm called

Figure 1a

Figure 1b

Figure 1. Two Grid Worlds where Expectation-Optimal and Minimax-Optimal Policies Differ

Q-hat-learning that are based on the minimax criterion. According to this criterion, a policy is optimal if it minimizes the worst-case return. The main contribution of this paper is to present new results for minimax-based tasks. The following two examples are intended to convince the reader that it is worthwhile to consider the minimax criterion as an alternative to the expected-value criterion.

In the stochastic domain of Figure 1a, the minimax-optimal policy differs from the expectation-optimal one. An agent (robot) has to navigate on the grid from location S (start) to G (goal). Dark squares represent obstacles; so does the perimeter of the grid. In each square, the agent has the four actions UP, DOWN, LEFT, and RIGHT available that take it to the corresponding adjacent square, unless the square is occupied by an obstacle, in which case the agent does not change location. Every state transition has an immediate cost of one except for state transitions from the goal state into itself, which have cost zero. The square with the arrow is a probabilistic one-way door. It can be passed through only in the direction of the arrow and behaves like an obstacle when approached from the opposite direction. If the agent is on the one-way door square, then, with probability p , it moves through the door, no matter which action it executes. With probability $1 - p$, however, the agent behaves as if it were on an empty square.

Consider the case where $p = 0.75$. When following the expectation-optimal policy π , the agent has to go RIGHT in square S and hence it moves to the one-way door square,

where it has to select RIGHT again. The agent reaches its goal in the next step only with probability 0.25 because with probability 0.75 it falls through the one-way door. In the latter case, it needs 18 additional steps to move around the vertical obstacle to the goal. Hence it needs $0.25 \cdot 2 + 0.75 \cdot 20 = 15.5$ steps on average to reach the goal from the start. The minimax-optimal policy suggests to go DOWN in square S and move around the horizontal obstacle. This always takes 16 steps.

This example demonstrates that expectation-optimal policies do not necessarily have the lowest return in most cases. If, for example, the above experiment is repeated a large number of times then the expectation-optimal policy needs 25% more cost units in 75% of all cases when compared to the minimax-optimal policy. The latter policy minimizes the worst-case return and hence produces risk-avoiding behavior: the agent avoids the risk of falling through the one-way door.

In the second example (Figure 1b), the vertical obstacle is eleven squares longer than in the previous example. Consider the case where $p = 1/3$. Again, the expectation-optimal policy tells the agent to go across the one-way door because this requires on average $\frac{2 \cdot 2}{3} + \frac{42 \cdot 1}{3} = 15\frac{1}{3} < 16$ steps. In this domain, the expectation-optimal policy has the lowest return (2 steps) in most cases, but with probability $1/3$ the agent has to follow a very long path around the vertical obstacle. If the agent does not have enough energy for this long path (42 steps), it would be better to use the minimax-optimal policy (16 steps).

This example demonstrates that the expected value criterion is not reliable, especially when solving tasks with resource constraints. Resource constraints cannot be addressed by giving the agent a big punishment (i.e., high immediate cost) when its accumulated cost has exceeded the given limit, because these reinforcement signals depend on the history of the agent's interaction with its environment and, consequently, do not satisfy the Markov property – unless the state representation is augmented with information about the accumulated cost. This, however, would increase the size of the state space to an extent that makes efficient learning impossible. Fortunately, tasks for which the total cost is not allowed to exceed a given limit can be solved using the minimax criterion, because the minimax-optimal policy is guaranteed to obey this constraint if this is possible.

Recently, additional minimax-based reinforcement learning algorithms have been published. The model-based parti-game algorithm (Moore & Atkeson, 1995) uses the minimax criterion for navigation tasks in conjunction with an adaptive state-action space representation. Using the minimax criterion has the advantage that it makes the planning algorithm fast, allows it to use a simple world model, and automatically detects if a planning problem does not have a “safe” solution (in which case the partgame algorithm changes the state-action space). Littman uses the minimax-criterion for Markov games (Littman, 1994) that have two kinds of uncertainties: the transition uncertainty (as commonly represented by MDPs) and, additionally, uncertainty introduced by an opponent who does not base his decisions on well defined stationary probabilities. To deal with the latter uncertainty, Littman uses the minimax criterion as is common in game theory.

Before we formalize and analyze the problem addressed in this article, we describe the theory behind Q -learning and Q -hat-learning in Sections 3 and 4, respectively. First,

however, we introduce our notation and give an overview of the commonly used RL definitions and theorems.

2. Basic Definitions and Theorems

We define the *set of admissible state-action pairs* to be

$$M := \{(i, a) \in S \times A : a \in A(i)\}$$

and call the elements of $\mathcal{R}^M$ state-action value functions or, synonymously, Q-functions. Assume that a function $Q \in \mathcal{R}^M$ and a state $i \in S$ are given. Then, we call an action

$$a \in A(i) \text{ greedy for } i \text{ and } Q \text{ iff } a = \arg \min_{b \in A(i)} Q(i, b).$$

Furthermore, we call a stationary policy

$$\pi \text{ greedy for } Q \text{ iff } \forall i \in S : \pi(i) = \arg \min_{a \in A(i)} Q(i, a).$$

Finally, we call a function $Q \in \mathcal{R}^M$ *greedy-equivalent* to another function $Q' \in \mathcal{R}^M$ iff

$$\{a \in A(i) : a \text{ is greedy for } i \text{ and } Q\} = \{a \in A(i) : a \text{ is greedy for } i \text{ and } Q'\}$$

for all $i \in S$. Let $\epsilon > 0$, $i \in S$, $a \in A(i)$, and $Q \in \mathcal{R}^M$. We say that a is ϵ -greedy with respect to i and Q iff $|Q(i, a) - \min_{b \in A(i)} Q(i, b)| \leq \epsilon$.

We define a partial order on Q-functions as follows: $Q \leq Q'$ iff $Q(i, a) \leq Q'(i, a)$ for all $i \in S$ and $a \in A(i)$. For a given real number r and a Q-function Q we write $Q + r$ and $r \cdot Q$ to denote the Q-functions Q' and Q'' that satisfy $Q'(i, a) = Q(i, a) + r$ and $Q''(i, a) = r \cdot Q(i, a)$, respectively, for all $i \in S$ and $a \in A(i)$. We write $Q \leq r$ iff $Q(i, a) \leq r$ for all $i \in S$ and $a \in A(i)$. We use the maximum norm $\|\cdot\|$ for Q-functions, i.e.,

$$\|Q\| = \max_{i \in S, a \in A(i)} |Q(i, a)|.$$

An element of $\mathcal{R}^S$ is called a (state) value function. For a given real number r and a value function V we write $V + r$ and $r \cdot V$ to denote the value function V' and V'' that satisfy $V'(i) = V(i) + r$ and $V''(i) = r \cdot V(i)$, respectively, for all $i \in S$. We write $V \leq r$ iff $V(i) \leq r$ for all $i \in S$. We use the maximum norm $\|\cdot\|$ for value functions, i.e.,

$$\|V\| = \max_{i \in S} |V(i)|.$$

For a given Q-function Q and stationary policy π we define the value functions $V_{Q, \pi}$ and V_Q to be

$$\forall i \in S : V_{Q, \pi}(i) = Q(i, \pi(i)) \text{ and } V_Q(i) = \min_{a \in A(i)} Q(i, a),$$

respectively.

In the following, we state six theorems that are used to prove the results of Section 5. The first three theorems follow immediately from the definitions of this section.

THEOREM 1 *Let Q be a Q -function, π a stationary policy, and $r \in \mathcal{R}$. Then, $V_{Q+r} = V_Q + r$ and $V_{Q+r, \pi} = V_{Q, \pi} + r$.*

THEOREM 2 *Let Q and Q' be Q -functions with $Q \leq Q'$ and let π be any stationary policy. Then, $V_Q \leq V_{Q'}$ and $V_{Q, \pi} \leq V_{Q', \pi}$.*

THEOREM 3 *Let π be a stationary policy and let Q be a Q -function. Then π is greedy for Q iff $V_{Q, \pi} = V_Q$.*

The following theorem is proven in (Williams & Baird, 1993):

THEOREM 4 *Let g_1 and g_2 be real-valued functions on a compact domain U . Then*

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

THEOREM 5 *Let g_1 and g_2 be real-valued functions on a compact domain U . Then*

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

Proof: Let $u_1 = \arg \min_{u \in U} g_1(u)$ and $u_2 = \arg \min_{u \in U} g_2(u)$. Consider first the case where $g_1(u_1) \geq g_2(u_2)$. Then

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

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

THEOREM 6 *For any stationary policy π and arbitrary Q -functions Q and Q' it holds that $\|V_{Q, \pi} - V_{Q', \pi}\| \leq \|Q - Q'\|$ and $\|V_Q - V_{Q'}\| \leq \|Q - Q'\|$.*

Proof: For any state i ,

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

Since this holds for all states, the first inequality follows. For the second inequality, consider an arbitrary state i . Applying Theorem 5 yields

$$\begin{aligned} |V_Q(i) - V_{Q'}(i)| &= \left| \min_{a \in A(i)} Q(i, a) - \min_{a \in A(i)} Q'(i, a) \right| \\ &\leq \max_{a \in A(i)} |Q(i, a) - Q'(i, a)| \leq \|Q - Q'\|, \end{aligned}$$

and the desired result follows. ■

3. Expectation-Based Markov Decision Tasks

For expectation-based Markov decision tasks, one uses the value function $V_\gamma^\pi \in \mathcal{R}^S$ to measure the performance of a given policy π for a given MDP. The value function is defined as follows:

$$\forall i \in S : V_\gamma^\pi(i) = E(R_\gamma^\pi \mid I_0^\pi = i),$$

i.e., $V_\gamma^\pi(i)$ is the return one expects if the agent starts in state i and uses policy π . A policy is called optimal if its value function is equal to the optimal value function $V_\gamma^* \in \mathcal{R}^S$ that is defined by

$$\forall i \in S : V_\gamma^*(i) = \inf_{\pi} V_\gamma^\pi(i).$$

3.1. Dynamic Programming Operators

A fundamental result from the theory of dynamic programming states that there always exists an optimal policy that is stationary. In dynamic programming, one approximates value functions of policies and optimal value functions by applying dynamic programming operators repeatedly. The dynamic programming operator $T_\gamma : \mathcal{R}^S \rightarrow \mathcal{R}^S$ is a mapping from value functions to value functions, defined by

$$\forall V \in \mathcal{R}^S \forall i \in S : T_\gamma V(i) = \min_{a \in A(i)} \left[R(i, a) + \gamma \cdot \sum_{j \in S} P_S(i, a, j) V(j) \right]$$

where $R(i, a)$ denotes the expected immediate cost for a given starting state i and executed action a .¹ For convenience we write $T_\gamma V(i)$ instead of $[T_\gamma(V)](i)$ and proceed similarly for other dynamic programming operators defined below. For a given stationary policy π , the dynamic programming operator $T_\gamma^\pi : \mathcal{R}^S \rightarrow \mathcal{R}^S$ is a mapping from value functions to value functions, defined by

$$\forall V \in \mathcal{R}^S \forall i \in S : T_\gamma^\pi V(i) = R(i, \pi(i)) + \gamma \cdot \sum_{j \in S} P_S(i, \pi(i), j) V(j).$$

A policy is called greedy for a value function V if

$$\forall i \in S : \pi(i) = \arg \min_{a \in A(i)} \left[R(i, a) + \gamma \cdot \sum_{j \in S} P_S(i, a, j) \cdot V(j) \right].$$

The following three theorems are known from the theory of dynamic programming (see, e.g., Bertsekas (1987)):

THEOREM 7 *Let $V_0 \in \mathcal{R}^S$ and $V_{k+1} = T_\gamma V_k$ for all $i \in S$ and $k \in \mathbb{N}$. Then $\lim_{k \rightarrow \infty} \|V_k - V_\gamma^*\| = 0$.*

THEOREM 8 *Let $\pi \in A^S$ be a stationary policy, $V_0 \in \mathcal{R}^S$, and $V_{k+1} = T_\gamma^\pi V_k$ for all $i \in S$ and $k \in \mathbb{N}$. Then $\lim_{k \rightarrow \infty} \|V_k - V_\gamma^\pi\| = 0$.*

THEOREM 9 *A stationary policy π is optimal iff it is greedy for V_γ^* .*

Theorem 7 yields an algorithm for computing the optimal value function, and Theorem 8 yields a similar algorithm for computing the value function of a given policy. Theorem 9 shows that optimal policies can be determined easily from a given optimal state value function, which is the reason why learning algorithms often approximate the optimal value function. To be able to compute a greedy policy for a given value function, one needs to know a world model, in this case the probabilities for the state transitions and the associated expected immediate costs.

We conclude this section by defining the dynamic programming operators B_γ and B_γ^π for stationary policies π . These operators are associated with Q-functions and the expected-value criterion, and will be used in Section 5. $B_\gamma : \mathcal{R}^M \rightarrow \mathcal{R}^M$ satisfies

$$B_\gamma Q(i, a) = R(i, a) + \gamma \cdot \sum_{j \in S} P_S(i, a, j) V_Q(j)$$

for any Q-function Q , state i , and action $a \in A(i)$. $B_\gamma^\pi : \mathcal{R}^M \rightarrow \mathcal{R}^M$ is defined by

$$B_\gamma^\pi Q(i, a) = R(i, a) + \gamma \cdot \sum_{j \in S} P_S(i, a, j) V_{Q, \pi}(j)$$

for any Q-function Q , state i , and action $a \in A(i)$.

In the next section, we introduce a reinforcement learning algorithm that learns a Q-function with the property that every policy that is greedy for this Q-function is optimal according to the expected value criterion. Greedy policies can be obtained from Q-functions even if there is no world model available.

3.2. Q-Learning

For a stationary policy π , we define the Q-function Q_γ^π by

$$Q_\gamma^\pi(i, a) = R(i, a) + \gamma \cdot \sum_{j \in S} P_S(i, a, j) \cdot V_\gamma^\pi(j)$$

for all $i \in S$ and $a \in A(i)$. The Q-learning algorithm (Watkins, 1989) approximates the optimal Q-function Q_γ^* that is defined by

$$Q_\gamma^*(i, a) = R(i, a) + \gamma \cdot \sum_{j \in S} P_S(i, a, j) \cdot V_\gamma^*(j).$$

From Theorem 9, we conclude that a stationary policy is optimal iff it is greedy for Q_γ^* . The Q-learning algorithm begins with an initial estimate Q_0 of Q_γ^* and improves this estimate as the results of individual actions become apparent. After each episode

t , the Q-value associated with the starting state i and executed action a is updated as follows:

$$Q_{t+1}(i, a) = \alpha_t(i, a) (r + \gamma V_{Q_t}(j)) + (1 - \alpha_t(i, a)) Q_t(i, a)$$

where j is the successor state reached and r is the immediate cost of the episode. The learning rates $\alpha_t(i, a)$ have to satisfy

$$\sum_{t=0}^{\infty} \alpha_t(i, a) = \infty; \quad \sum_{t=0}^{\infty} \alpha_t(i, a)^2 < \infty.$$

In (Watkins & Dayan, 1992) and recently in (Tsitsiklis, 1994) it is proven that the Q-learning algorithm converges to Q_γ^* if every action that is admissible in a state is executed infinitely often in that state.

4. Minimax-Based Markov Decision Tasks

The theory behind minimax-based Markov decision tasks is very similar to the one behind expectation-based tasks, as will become apparent in the following. We therefore use the same notation where appropriate, which also simplifies our presentation of a unified analysis of both kinds of Markov decision tasks in a later section. From the context, it is always obvious which kind of task we are referring to.

The definitions in this section do not depend on the earlier definitions for expectation-based Markov decision tasks, which avoids cyclic definitions. For minimax-based Markov decision tasks, we use the value function $V_\gamma^\pi \in \mathcal{R}^S$ to measure the performance of a given policy π for a given MDP. The value function is defined as follows:

$$\forall i \in S : V_\gamma^\pi(i) = \sup \{r \in \mathcal{R} : P(R_\gamma^\pi > r \mid I_0^\pi = i) > 0\},$$

i.e., $V_\gamma^\pi(i)$ is the worst-case return that can possibly occur if the agent starts in state i and uses policy π . A policy is called optimal if its value function is equal to the optimal value function $V_\gamma^* \in \mathcal{R}^S$ that is defined by

$$\forall i \in S : V_\gamma^*(i) = \inf_{\pi} V_\gamma^\pi(i).$$

4.1. Dynamic Programming Operators

From (Heger, 1994b) we obtain the result that there always exists an optimal policy that is stationary. Let

$$c(i, a, j) = \sup \{r \in C : P_C(i, a, j, r) > 0\}$$

denote the worst-case immediate cost that can be obtained for the state transition from i to j under action a , and let

$$N(i, a) = \{j \in S : P_S(i, a, j) > 0\}$$

be the set of possible successor states when executing action a in state i . The dynamic programming operator $T_\gamma : \mathcal{R}^S \rightarrow \mathcal{R}^S$ is a mapping from value functions to value functions, defined by

$$\forall V \in \mathcal{R}^S \forall i \in S : T_\gamma V(i) = \min_{a \in A(i)} \max_{j \in N(i,a)} [c(i, a, j) + \gamma \cdot V(j)]. \quad (1)$$

For a given stationary policy π , the dynamic programming operator $T_\gamma^\pi : \mathcal{R}^S \rightarrow \mathcal{R}^S$ is a mapping from value functions to value functions, defined by

$$\forall V \in \mathcal{R}^S \forall i \in S : T_\gamma^\pi V(i) = \max_{j \in N(i, \pi(i))} [c(i, \pi(i), j) + \gamma \cdot V(j)]. \quad (2)$$

A policy is called greedy for a value function V if

$$\forall i \in S : \pi(i) = \arg \min_{a \in A(i)} \left[\max_{j \in N(i,a)} (c(i, a, j) + \gamma \cdot V(j)) \right].$$

The following three theorems are proven in (Heger, 1994b):

THEOREM 10 *Let $\pi \in A^S$ be a stationary policy, $V_0 \in \mathcal{R}^S$, and $V_{k+1} = T_\gamma V_k$ for all $i \in S$ and $k \in \mathbb{N}$. Then $\lim_{k \rightarrow \infty} \|V_k - V_\gamma^*\| = 0$.*

THEOREM 11 *Let $\pi \in A^S$ be a stationary policy, $V_0 \in \mathcal{R}^S$, and $V_{k+1} = T_\gamma^\pi V_k$ for all $i \in S$ and $k \in \mathbb{N}$. Then $\lim_{k \rightarrow \infty} \|V_k - V_\gamma^\pi\| = 0$.*

THEOREM 12 *A stationary policy π is optimal iff it is greedy for V_γ^* .*

The first two theorems yield algorithms for computing the optimal value function and the value function of a given policy, respectively. Theorem 12 shows that optimal policies can be determined easily from a given optimal value function. To be able to compute a greedy policy for a given value function, one needs to know a world model, in this case the sets of possible successor states for all state-action pairs and the worst immediate costs which the action executions can result in. Note that this world model is, in general, much simpler than the corresponding world model needed for expectation-based Markov decision tasks.

We conclude this section by defining the dynamic programming operators B_γ and B_γ^π for stationary policies π . These operators are associated with Q-functions and the minimax criterion, and will be used in Section 5. $B_\gamma : \mathcal{R}^M \rightarrow \mathcal{R}^M$ satisfies

$$B_\gamma Q(i, a) = \max_{j \in N(i,a)} [c(i, a, j) + \gamma \cdot V_Q(j)]$$

for any Q-function Q , state i , and action $a \in A(i)$. $B_\gamma^\pi : \mathcal{R}^M \rightarrow \mathcal{R}^M$ is defined by

$$B_\gamma^\pi Q(i, a) = \max_{j \in N(i, \pi(a))} [c(i, a, j) + \gamma \cdot V_{Q, \pi}(j)]$$

for any Q-function Q , state i , and action $a \in A(i)$.

In the next section, we introduce a reinforcement learning algorithm that learns a Q-function with the property that every policy that is greedy for this Q-function is optimal according to the minimax criterion. Its advantage is, as in the case of expectation-based Markov decision tasks, that one does not need to know a world model to compute a policy that is greedy for a given Q-function.

4.2. Q-Hat-Learning

For a stationary policy π , we define the Q-function Q_γ^π by

$$Q_\gamma^\pi(i, a) = \max_{j \in N(i, a)} [c(i, a, j) + \gamma \cdot V_\gamma^\pi(j)]$$

for all $i \in S$ and $a \in A(i)$. The Q-hat-learning algorithm (Heger, 1994a) approximates the optimal Q-function Q_γ^* that is defined by

$$Q_\gamma^*(i, a) = \max_{j \in N(i, a)} [c(i, a, j) + \gamma \cdot V_\gamma^*(j)] .$$

From Theorem 12, we conclude that a stationary policy is optimal iff it is greedy for Q_γ^* . The Q-hat-learning algorithm begins with an initial estimate $Q_0 \leq Q_\gamma^*$ and improves this estimate as the results of individual actions become apparent. After each episode t , the Q-value associated with the starting state i and executed action a is updated as follows:

$$Q_{t+1}(i, a) = \max \{Q_t(i, a), r + \gamma V_{Q_t}(j)\}$$

where j is the successor state reached and r is the immediate cost of the episode. Note that the Q-hat-learning algorithm has no need for a learning rate, and that the Q-values are monotonely increasing in time. In (Heger, 1995) it is proven that the Q-hat-learning algorithm converges to Q_γ^* if every action that is admissible in a state is executed infinitely often in that state.

An advantage of the Q-hat-learning algorithm is that the exploration-exploitation problem (e.g., Thrun, 1992) for Q-hat-learning is not as severe as the one for Q-learning: Let $\epsilon > 0$ be a constant and assume that every state becomes infinitely often a starting state. Assume further that the agent always selects an action from the set of actions that are ϵ -greedy with respect to its current Q-function and starting state, and that it selects an action from this set with uniform probability. Then, the Q-hat-learning algorithm converges, with probability one, to a Q-function $\hat{Q}$ that is greedy-equivalent to Q_γ^* , and $\hat{Q}(i, a) = Q_\gamma^*(i, a)$ with probability one for every state $i \in S$ and action $a \in A(i)$ that is greedy for i and Q_γ^* .

5. The Loss from Imperfect Value Functions

In the previous two sections, we have seen that expectation-based and minimax-based tasks have in common that optimal policies can be determined by finding policies that are greedy for the optimal value function. This is the reason why many dynamic programming approaches, such as the famous Q-learning algorithm and its counterpart for the minimax criterion, the Q-hat-learning algorithm, are only concerned with finding an optimal value function.

The question remains to which extent the performance of the agent differs from its optimal performance if it uses a policy that is greedy for an *approximation* of the optimal value function instead of a policy that is greedy for the optimal value function itself;

see for example (Singh & Yee, 1994) and (Williams & Baird, 1993). Knowing the magnitude of this loss in performance is important for at least two reasons: First, dynamic programming and RL algorithms iteratively improve an approximation of the optimal value function. The value function obtained after a finite number of iterations is, in general, only an approximation of the optimal value function. Second, one cannot use lookup tables to represent value functions in complex domains, because there are too many states. This makes it necessary to use parametric function approximators, such as neural networks, that have the ability to generalize. Since function approximators cannot guarantee that they generalize correctly, they can only be expected to approximate the optimal value function.

In practice, one can therefore only expect to obtain approximations of optimal value functions. If small deviations of these approximations from the optimal value functions resulted in an arbitrarily bad performance of the agent, this would raise significant concerns about the use of function approximators in dynamic programming-based learning.

5.1. Results for Value Functions

The maximum norm distance was used in (Singh & Yee, 1994) to measure the approximation quality of a value function; in (Williams & Baird, 1993), the so-called Bellman error magnitude was used for the same purpose. In both cases, the loss in performance turned out to be at most proportional to the approximation error for expectation-based Markov decision tasks.

We generalize this work by giving an abstract analysis that holds for both expectation-based tasks and minimax-based tasks. We only use properties that dynamic programming operators of both tasks have in common. In the following, we present these common features as “assumptions” and defer their proof to Appendix A. Based on these assumptions, we present in this and the following section our results about the loss from imperfect value functions and Q-functions, respectively. The significance of these results will be discussed in the conclusion of this article, and the corresponding proofs can be found in Appendix B.

Assume V_γ^* and V_γ^π (for some stationary policy π) are value functions, and T_γ and $T_\gamma^\pi : \mathcal{R}^S \rightarrow \mathcal{R}^S$ are operators. These value functions and operators have to satisfy the following eight assumptions.

The first assumption guarantees that the value function V_γ^π (V_γ^*) is the only fixed point of the operator T_γ^π (T_γ).

Assumption 1 *Let π be a stationary policy and V be a value function. Then $T_\gamma^\pi V = V$ iff $V = V_\gamma^\pi$, and $T_\gamma V = V$ iff $V = V_\gamma^*$.*

Assume that we apply operator T_γ or T_γ^π to a value function that is the sum of a real value r and a value function V . The next assumption states that we obtain $\gamma \cdot r$ plus the value function that is obtained by applying the same operator to V .

Assumption 2 *Let π be a stationary policy, V be a value function, and $r \in \mathcal{R}$. Then $T_\gamma(V + r) = T_\gamma V + \gamma \cdot r$ and $T_\gamma^\pi(V + r) = T_\gamma^\pi V + \gamma \cdot r$.*

Third, we assume that both operators maintain the partial relation $\leq$ that was defined earlier for value functions.

Assumption 3 *Let U and V be value functions with $U \leq V$ and π be a stationary policy. Then $T_\gamma U \leq T_\gamma V$ and $T_\gamma^\pi U \leq T_\gamma^\pi V$.*

Furthermore, we assume that the maximum norm distance between the images of any two value functions U and V is at most γ times the maximum norm distance between U and V , no matter which of the two operators is used.

Assumption 4 *Let π be a stationary policy and U and V be value functions. Then $\|T_\gamma U - T_\gamma V\| \leq \gamma \cdot \|U - V\|$ and $\|T_\gamma^\pi U - T_\gamma^\pi V\| \leq \gamma \cdot \|U - V\|$.*

The next assumption defines greedy policies in an abstract way that does not depend on details of either expectation-based or minimax-based decision tasks. It also shows the fundamental relationship between the operators T_γ^π and T_γ .

Assumption 5 *Let π be any stationary policy and V be any value function. Then π is greedy for V iff $T_\gamma^\pi V = T_\gamma V$.*

The following assumption shows the fundamental relationship between the value functions V_γ^* and V_γ^π .

Assumption 6 *Let i be any state and Π_S be the set of stationary policies. Then $V_\gamma^*(i) = \min_{\pi \in \Pi_S} V_\gamma^\pi(i)$.*

For deterministic MDPs, the starting state i and the executed action a determine uniquely the successor state j_i^a reached and the immediate cost c_i^a obtained. We assume that for deterministic MDPs, the operators T_γ and T_γ^π are defined in the following way.

Assumption 7 *Let π be a stationary policy, i be a state, and V be a value function. Then $T_\gamma V(i) = \min_{a \in A(i)} [c_i^a + \gamma \cdot V(j_i^a)]$ and $T_\gamma^\pi V(i) = c_i^{\pi(i)} + \gamma \cdot V(j_i^{\pi(i)})$ if the MDP is deterministic.*

Consider a deterministic MDP, any stationary policy π , and a given starting state i at time $t = 0$. In this case the return, call it $r_\gamma^\pi(i)$, is uniquely determined. The eighth and final assumption states that V_γ^π is equal to this value.²

Assumption 8 *$V_\gamma^\pi(i) = r_\gamma^\pi(i)$ for any deterministic Markov decision process, stationary policy π , and state i .*

Before we can analyze the loss in performance due to imperfect value functions we have to define what we precisely mean by “imperfect.” We measure the distance between a value function V and the optimal value function V_γ^* in two different ways:

$$d_M(V) := \|V - V_\gamma^*\| = \max_{i \in S} |V(i) - V_\gamma^*(i)|; \quad (3)$$

$$d_B(V) := \|T_\gamma V - V\| = \max_{i \in S} |T_\gamma V(i) - V(i)|. \quad (4)$$

We call $d_M(V)$ the *maximum norm distance* between V and V_γ^* and, following Williams and Baird, $d_B(V)$ the *Bellman error magnitude*. From its definition it is not obvious that the Bellman error magnitude is a reasonable measure for the distance between V and V_γ^* . The next two theorems show that this is indeed the case.

THEOREM 13 *Let V be any value function and π be any stationary policy. Then*

$$\begin{aligned} (1 - \gamma) \cdot \|V - V_\gamma^*\| &\leq \|V - T_\gamma V\| \leq (1 + \gamma) \cdot \|V - V_\gamma^*\|; \\ (1 - \gamma) \cdot \|V - V_\gamma^\pi\| &\leq \|V - T_\gamma^\pi V\| \leq (1 + \gamma) \cdot \|V - V_\gamma^\pi\|. \end{aligned}$$

THEOREM 14 *Let V be a value function that satisfies $V \leq V_\gamma^*$ and π be any stationary policy. Then*

$$\begin{aligned} (1 - \gamma) \cdot \|V - V_\gamma^*\| &\leq \|V - T_\gamma V\| \leq \|V - V_\gamma^*\|; \\ (1 - \gamma) \cdot \|V - V_\gamma^\pi\| &\leq \|V - T_\gamma^\pi V\| \leq \|V - V_\gamma^\pi\|. \end{aligned}$$

Theorem 13 ensures that $(1 - \gamma) \cdot d_M(V) \leq d_B(V) \leq (1 + \gamma) \cdot d_M(V)$. This tight relationship between the Bellman error magnitude and the maximum norm distance justifies the use of the Bellman error magnitude as a measure for the distance between V and V_γ^* . Its advantage over the maximum norm distance is that it can be computed easily even if V_γ^* is unknown.

We measure the loss in performance due to a policy π by $\|V_\gamma^\pi - V_\gamma^*\|$. The following two theorems guarantee that the loss in performance is small if dynamic programming based learning approaches are used that satisfy the assumptions above, good approximations of the optimal value functions are achieved, and then a greedy policy is followed – provided that the discount factor γ is not too close to 1.0.

THEOREM 15 *Let V be a value function and π be a stationary policy that is greedy for V . Then $\|V_\gamma^\pi - V_\gamma^*\| \leq \frac{2\gamma}{1-\gamma} \cdot d_B(V)$. Furthermore, this bound is tight, i.e., there exists an example with $\|V_\gamma^\pi - V_\gamma^*\| = \frac{2\gamma}{1-\gamma} \cdot d_B(V)$.*

THEOREM 16 *Let V be a value function and π be a stationary policy that is greedy for V . Then $\|V_\gamma^\pi - V_\gamma^*\| \leq \frac{2\gamma}{1-\gamma} \cdot d_M(V)$. Furthermore, this bound is tight.*

The next theorem and corollary guarantee that the bound for the loss in performance can be reduced by a factor of 2 if the approximation V underestimates V_γ^* .

THEOREM 17 *Let V be a value function with $V \leq V_\gamma^*$ and π be a stationary policy that is greedy for V . Then $\|V_\gamma^\pi - V_\gamma^*\| \leq \frac{\gamma}{1-\gamma} \cdot d_B(V)$. Furthermore, this bound is tight.*

COROLLARY 1 *Let V be a value function with $V \leq V_\gamma^*$ and π be a stationary policy that is greedy for V . Then $\|V_\gamma^\pi - V_\gamma^*\| \leq \frac{\gamma}{1-\gamma} \cdot d_M(V)$. Furthermore, this bound is tight.*

It might not always be easy to verify that $V \leq V_\gamma^*$ if V_γ^* is unknown, but the following two corollaries provide a sufficient condition for $V \leq V_\gamma^*$ that depends only on the dynamic programming operator T_γ and is easy to verify.

COROLLARY 2 *Let V be a value function and π be a greedy policy for V . If $V \leq T_\gamma V$, then $\|V_\gamma^\pi - V_\gamma^*\| \leq \frac{\gamma}{1-\gamma} \cdot d_B(V)$. Furthermore, this bound is tight.*

COROLLARY 3 *Let V be a value function and π be a greedy policy for V . If $V \leq T_\gamma V$, then $\|V_\gamma^\pi - V_\gamma^*\| \leq \frac{\gamma}{1-\gamma} \cdot d_M(V)$. Furthermore, this bound is tight.*

5.2. Results for Q-Functions

This section proceeds similarly to the last section, except that our results now apply to Q-functions instead of value functions: We first present features of dynamic programming operators that hold for both expectation-based and minimax-based Markov decision tasks and then derive theorems based on these assumptions.

Assume that V_γ^* and V_γ^π (for some stationary policy π) are value functions, Q_γ^* and Q_γ^π are Q-functions, and $B_\gamma : \mathcal{R}^M \rightarrow \mathcal{R}^M$ and $B_\gamma^\pi : \mathcal{R}^M \rightarrow \mathcal{R}^M$ are operators. These value functions, Q-functions, and operators have to satisfy the following assumptions in addition to the eight assumptions stated in the previous section. We show in Appendix A that all of these assumptions are satisfied for both expectation-based and minimax-based Markov decision tasks.

Assumption 9 relates the value functions V_γ^* and V_γ^π to their corresponding Q-functions Q_γ^* and Q_γ^π , respectively.

Assumption 9 *Let π be a stationary policy. Then $V_{Q_\gamma^*} = V_\gamma^*$ and $V_{Q_\gamma^\pi, \pi} = V_\gamma^\pi$.*

Assumption 10 guarantees that the Q-function Q_γ^π (Q_γ^*) is the only fixed point of the operator B_γ^π (B_γ).

Assumption 10 *Let π be a stationary policy and Q be a Q-function. Then $B_\gamma Q = Q$ iff $Q = Q_\gamma^*$, and $B_\gamma^\pi Q = Q$ iff $Q = Q_\gamma^\pi$.*

Assume that we apply operator B_γ or B_γ^π to a Q-function that is the sum of a real value r and a Q-function Q . Assumption 11 states that we obtain $\gamma \cdot r$ plus the Q-function that is obtained by applying the same operator to Q .

Assumption 11 *Let π be a stationary policy, Q be a Q-function, and $r \in \mathcal{R}$. Then $B_\gamma(Q + r) = B_\gamma Q + \gamma \cdot r$ and $B_\gamma^\pi(Q + r) = B_\gamma^\pi Q + \gamma \cdot r$.*

Assumption 12 states that both operators maintain the partial relation $\leq$ for Q-functions.

Assumption 12 *Let Q and Q' be Q-functions with $Q \leq Q'$ and π be a stationary policy. Then $B_\gamma Q \leq B_\gamma Q'$ and $B_\gamma^\pi Q \leq B_\gamma^\pi Q'$.*

We assume that the maximum norm distance between the images of any two Q-functions Q and Q' is at most γ times the maximum norm distance between Q and Q' , no matter which of the two operators is used.

Assumption 13 *Let π be a stationary policy and Q and Q' be Q-functions. Then $\|B_\gamma Q - B_\gamma Q'\| \leq \gamma \cdot \|Q - Q'\|$ and $\|B_\gamma^\pi Q - B_\gamma^\pi Q'\| \leq \gamma \cdot \|Q - Q'\|$.*

Assumption 14 shows the fundamental relationship between the operators B_γ^π and B_γ .

Assumption 14 Let Q be a Q -function and π be a stationary policy that is greedy for Q . Then $B_\gamma Q = B_\gamma^\pi Q$.

Assumption 15 shows the fundamental relationship between the Q -functions Q_γ^* and Q_γ^π .

Assumption 15 Let i be a state, $a \in A(i)$, and Π_S be the set of stationary policies. Then $Q_\gamma^*(i, a) = \min_{\pi \in \Pi_S} Q_\gamma^\pi(i, a)$.

The following assumption gives the definitions of the operators B_γ and B_γ^π for deterministic MDPs.

Assumption 16 Let π be a stationary policy, i be a state, $a \in A(i)$ be an action, and Q be a Q -function. Then $B_\gamma Q(i, a) = c_i^a + \gamma \cdot V_Q(j_i^a)$ and similarly, $B_\gamma^\pi Q(i, a) = c_i^a + \gamma \cdot V_{Q, \pi}(j_i^a)$ if the MDP is deterministic.

The final assumption relates the definition of an optimal policy to the Q -function Q_γ^* .

Assumption 17 A stationary policy is optimal iff it is greedy for Q_γ^* .

Before we can analyze the loss in performance due to imperfect Q -functions we have to define what we precisely mean by “imperfect.” We measure the distance between a Q -function Q and the optimal Q -function Q_γ^* in two different ways:

$$d_M(Q) := \|Q - Q_\gamma^*\| = \max_{i \in S, a \in A(i)} |Q(i, a) - Q_\gamma^*(i, a)|; \quad (5)$$

$$d_B(Q) := \|B_\gamma Q - Q\| = \max_{i \in S, a \in A(i)} |B_\gamma Q(i, a) - Q(i, a)|. \quad (6)$$

We call $d_M(Q)$ the *maximum norm distance* between Q and Q_γ^* and, following Williams and Baird (Williams & Baird, 1993), $d_B(Q)$ the *Bellman error magnitude*.

THEOREM 18 Let Q be a Q -function and π be a stationary policy. Then

$$\begin{aligned} (1 - \gamma) \cdot \|Q - Q_\gamma^*\| &\leq \|Q - B_\gamma Q\| \leq (1 + \gamma) \cdot \|Q - Q_\gamma^*\|; \\ (1 - \gamma) \cdot \|Q - Q_\gamma^\pi\| &\leq \|Q - B_\gamma^\pi Q\| \leq (1 + \gamma) \cdot \|Q - Q_\gamma^\pi\|. \end{aligned}$$

THEOREM 19 Let Q be a Q -function with $Q \leq Q_\gamma^*$ and π be a stationary policy. Then

$$\begin{aligned} (1 - \gamma) \cdot \|Q - Q_\gamma^*\| &\leq \|Q - B_\gamma Q\| \leq \|Q - Q_\gamma^*\|; \\ (1 - \gamma) \cdot \|Q - Q_\gamma^\pi\| &\leq \|Q - B_\gamma^\pi Q\| \leq \|Q - Q_\gamma^\pi\|. \end{aligned}$$

Theorem 18 ensures that $(1 - \gamma) \cdot d_M(Q) \leq d_B(Q) \leq (1 + \gamma) \cdot d_M(Q)$. This relationship justifies to use the Bellman error magnitude as a measure for the distance between Q and Q_γ^* . Its advantage over the maximum norm distance is that it can be computed easily even if Q_γ^* is unknown.

The following two theorems guarantee that the loss in performance is small if dynamic programming based learning approaches are used that satisfy the assumptions above, have achieved good approximations of the optimal Q -function, and then follow a greedy policy – provided that the discount factor γ is not too close to 1.0.

THEOREM 20 *Let Q be a Q -function and π be a stationary policy that is greedy for Q . Then $\|V_\gamma^\pi - V_\gamma^*\| \leq \frac{2}{1-\gamma} \cdot d_B(Q)$. Furthermore, this bound is tight.*

THEOREM 21 *Let Q be a Q -function and π be a stationary policy that is greedy for Q . Then $\|V_\gamma^\pi - V_\gamma^*\| \leq \frac{2}{1-\gamma} \cdot d_M(Q)$. Furthermore, this bound is tight.*

The next theorem and corollary guarantee that the bound for the loss in performance can be reduced by a factor of 2 if the approximation Q underestimates Q_γ^* . These results apply especially to Q -hat-learning because this algorithm operates with underestimations of the minimax-optimal Q -function.

THEOREM 22 *Let Q be a Q -function with $Q \leq Q_\gamma^*$ and π be a stationary policy that is greedy for Q . Then $\|V_\gamma^\pi - V_\gamma^*\| \leq \frac{1}{1-\gamma} \cdot d_B(Q)$. Furthermore, this bound is tight.*

COROLLARY 4 *Let Q be a Q -function with $Q \leq Q_\gamma^*$ and π be a stationary policy that is greedy for Q . Then $\|V_\gamma^\pi - V_\gamma^*\| \leq \frac{1}{1-\gamma} \cdot d_M(Q)$. Furthermore, this bound is tight.*

It might not always be easy to verify that $Q \leq Q_\gamma^*$ if Q_γ^* is unknown, but the following two corollaries provide a sufficient condition for $Q \leq Q_\gamma^*$ that depends only on the dynamic programming operator B_γ and is easy to verify.

COROLLARY 5 *Let Q be a Q -function with $Q \leq B_\gamma Q$ and π be a stationary policy that is greedy for Q . Then $\|V_\gamma^\pi - V_\gamma^*\| \leq \frac{1}{1-\gamma} \cdot d_B(Q)$. Furthermore, this bound is tight.*

COROLLARY 6 *Let Q be a Q -function with $Q \leq B_\gamma Q$ and π be a stationary policy that is greedy for Q . Then $\|V_\gamma^\pi - V_\gamma^*\| \leq \frac{1}{1-\gamma} \cdot d_M(Q)$. Furthermore, this bound is tight.*

The following theorem and corollary show that policies that are greedy for Q -functions are optimal if the Q -functions approximate the optimal Q -function closely.

THEOREM 23 *There exists an $\epsilon > 0$ with the following property: Every stationary policy is optimal if it is greedy for a Q -function Q with $d_M(Q) < \epsilon$.*

COROLLARY 7 *There exists an $\epsilon > 0$ with the following property: Every stationary policy is optimal if it is greedy for a Q -function Q with $d_B(Q) < \epsilon$.*

However, Q -functions that do not approximate the optimal Q -function well do not necessarily prevent one from obtaining optimal policies, as the following two results show.

THEOREM 24 *For every real number τ there exists a Q -function Q with $d_M(Q) > \tau$ that has the following property: Every stationary policy that is greedy for Q is optimal.*

COROLLARY 8 *For every real number τ there exists a Q -function Q with $d_B(Q) > \tau$ that has the following property: Every stationary policy that is greedy for Q is optimal.*

6. Conclusions

In this paper, we have studied both minimax-based and expectation-based Markov decision tasks and quantified how much performance is lost if one uses value functions that approximate the optimal value function instead of the optimal value function itself. In particular, we described the properties that both Markov decision tasks have in common (namely Assumptions 1 to 17) and based our analysis solely on these properties. Using this framework, we were able to transfer the results of (Williams & Baird, 1993) and (Singh & Yee, 1994) to minimax-based Markov decision tasks. To the best of our knowledge, our analysis is the first one that studies approximate value functions for minimax-based Markov decision tasks and the resulting loss in performance.

Our main result is that this loss in performance is at most proportional to the approximation error (Theorems 15, 16, 20 and 21). The factor of proportionality, however, depends on the discount factor γ and approaches infinity if γ approaches one. Unfortunately, there are some Markov decision tasks with delayed reinforcement for which it is not reliable to choose γ freely (Schwartz, 1993; see also McDonald & Hingson, 1994).

We obtained previously unknown results that hold for both the expectation-based tasks and the minimax-based tasks. The right hand side inequalities of the Theorems 13, 14, 18, and 19 we proved show that the Bellman error magnitude (used in (Williams & Baird, 1993) in order to measure the quality of approximation) is more related to the maximum norm distance (used in (Singh & Yee, 1994)) than was previously known. The results were useful to derive new bounds. Furthermore, on the one hand, we showed that the bounds given in (Singh & Yee, 1994) are tight, i.e., there are examples where these bounds are attained (Theorem 16 and 21). On the other hand if one assumes $V \leq V_\gamma^*$ or $Q \leq Q_\gamma^*$ then the corresponding bounds can be reduced by a factor of two. These new bounds were also proven to be tight (Corollaries 1 and 4).

In the $\hat{Q}$ -learning algorithm, the Q -function Q always satisfies $Q \leq Q_\gamma^*$ with probability one. This is generally not true in Q -learning. Hence the results of Theorem 22 and Corollary 4 are especially important in minimax-based tasks.

We showed that small non-zero approximation errors in Q -functions are already sufficient to obtain optimal policies (Theorem 23 and Corollary 7). As a consequence, every algorithm that iteratively produces a sequence of Q -functions that converges in infinity to the optimal Q -function, is able to determine optimal policies after a finite time. On the other hand we also showed that small approximation errors are not necessary to obtain optimal policies (Theorem 24 and Corollary 8).

It is still an open problem why expectation-based and minimax-based Markov decision tasks share so many properties. Another related open problem is whether there are other types of Markov decision tasks that satisfy our assumptions. Answers to these questions would probably lead to a better understanding of the existing reinforcement learning algorithms and might lead to the development of algorithms that use decision criteria that are different from the ones we use today.

We have seen that the loss in performance is small when acting greedily on imperfect value functions if the value functions approximate the optimal value function well. A related issue is whether the reinforcement learning algorithms that were designed to be

used in conjunction with lookup tables work equally well with function approximators. Theoretical results can be found in (Thrun & Schwartz, 1993), (Baird, 1995), and (Gordon, 1995). Practical results in using RL algorithms with function approximators are published in (Tesauro, 1992), (Lin, 1993), (Crites & Barto, 1995), (Boyan & Moore, 1995), (Zhang & Dietterich, 1995), and (Sutton, 1995). RL algorithms that are specifically designed to be used with function approximators were recently published in (Baird, 1995) and (Boyan & Moore 1995).

Acknowledgments

Many special thanks to Sven Koenig for his substantial help with the final editing of this article.

Appendix A

Proofs of the Assumptions

It is easy to see that the seventeen assumptions hold for expectation-based Markov decision tasks. They are either standard results from dynamic programming (see e.g., Bertsekas, 1987; Ross, 1970) or follow immediately from Section 3. Assumptions 9 and 10 were proven in (Williams & Baird, 1993).

The seventeen assumptions also hold for minimax-based Markov decision tasks. Assumption 1 was proven in (Heger, 1994b). We prove Assumptions 4, 9, 10, and 13 in the following. The remaining eleven assumptions then follow immediately from the results of Section 4.

Proof of Assumption 4 for Minimax-Based Markov Decision Tasks:

Assume $T_\gamma U(i^*) \geq T_\gamma V(i^*)$ for $i^* = \arg \max_{i \in S} |T_\gamma U(i) - T_\gamma V(i)|$, and let $a^* = \arg \min_{a \in A(i^*)} [\max_{j \in N(i^*, a)} [c(i^*, a, j) + \gamma V(j)]]$. Then, according to Theorem 4,

$$\begin{aligned}
 \|T_\gamma U - T_\gamma V\| &= |T_\gamma U(i^*) - T_\gamma V(i^*)| \\
 &= T_\gamma U(i^*) - T_\gamma V(i^*) \\
 &= \min_{a \in A(i^*)} \max_{j \in N(i^*, a)} [c(i^*, a, j) + \gamma U(j)] \\
 &\quad - \min_{a \in A(i^*)} \max_{j \in N(i^*, a)} [c(i^*, a, j) + \gamma V(j)] \\
 &\leq \max_{j \in N(i^*, a^*)} [c(i^*, a^*, j) + \gamma U(j)] \\
 &\quad - \max_{j \in N(i^*, a^*)} [c(i^*, a^*, j) + \gamma V(j)] \\
 &\leq \max_{j \in N(i^*, a^*)} |c(i^*, a^*, j) + \gamma U(j) - (c(i^*, a^*, j) + \gamma V(j))| \\
 &= \gamma \cdot \max_{j \in N(i^*, a^*)} |U(j) - V(j)| \leq \gamma \cdot \|U - V\|.
 \end{aligned}$$

A symmetrical arguments holds if $T_\gamma V(i^*) \geq T_\gamma U(i^*)$. This proves the first inequality.

To prove the second inequality, first consider the special case where $A(i) = \{\pi(i)\}$ for all $i \in S$. In this case, $T_\gamma = T_\gamma^\pi$ (see (1) and (2) at page 206) and the second inequality follows from the first one. To show that the second inequality holds in general, notice that it does not depend on any assumptions about the sets $A(i)$ other than $\pi(i) \in A(i)$, which is trivially true. ■

Proof of Assumption 9 for Minimax-Based Markov Decision Tasks: According to Assumption 1, which was proven in (Heger 1994b), it holds for any state i that

$$\begin{aligned} V_{Q_\gamma^*}(i) &= \min_{a \in A(i)} Q_\gamma^*(i, a) = \min_{a \in A(i)} \max_{j \in N(i, a)} [c(i, a, j) + \gamma \cdot V_\gamma^*(j)] \\ &= T_\gamma V_\gamma^*(i) = V_\gamma^*(i) \end{aligned}$$

which proves the first equation. Similarly, it follows from Assumption 1 that

$$\begin{aligned} V_{Q_\gamma^\pi, \pi}(i) &= Q_\gamma^\pi(i, \pi(i)) = \max_{j \in N(i, \pi(i))} [c(i, \pi(i), j) + \gamma \cdot V_\gamma^\pi(j)] \\ &= T_\gamma^\pi V_\gamma^\pi(i) = V_\gamma^\pi(i) \end{aligned}$$

which proves the second equation. ■

Proof of Assumption 10 for Minimax-Based Markov Decision Tasks: According to Assumption 9, it holds for any state i and action $a \in A(i)$ that

$$\begin{aligned} B_\gamma Q_\gamma^*(i, a) &= \max_{j \in N(i, a)} [c(i, a, j) + \gamma \cdot V_{Q_\gamma^*}(j)] \\ &= \max_{j \in N(i, a)} [c(i, a, j) + \gamma \cdot V_\gamma^*(j)] = Q_\gamma^*(i, a) \end{aligned}$$

which proves the first equation. Similarly, it follows from Assumption 9 that

$$\begin{aligned} B_\gamma^\pi Q_\gamma^\pi(i, a) &= \max_{j \in N(i, a)} [c(i, a, j) + \gamma \cdot V_{Q_\gamma^\pi, \pi}(j)] \\ &= \max_{j \in N(i, a)} [c(i, a, j) + \gamma \cdot V_\gamma^\pi(j)] = Q_\gamma^\pi(i, a) \end{aligned}$$

which proves the second equation. ■

Proof of Assumption 13 for Minimax-Based Markov Decision Tasks: Let $i \in S$ and $a \in A(i)$. From Theorems 4 and 6 we can conclude

$$\begin{aligned} &\|B_\gamma^\pi Q - B_\gamma^\pi Q'\| \\ &\leq |B_\gamma^\pi Q(i, a) - B_\gamma^\pi Q'(i, a)| \\ &= \left| \max_{j \in N(i, a)} [c(i, a, j) + \gamma \cdot V_{Q, \pi}(j)] - \max_{j \in N(i, a)} [c(i, a, j) + \gamma \cdot V_{Q', \pi}(j)] \right| \end{aligned}$$

$$\begin{aligned}
&\leq \max_{j \in N(i,a)} |c(i,a,j) + \gamma \cdot V_{Q,\pi}(j) - (c(i,a,j) + \gamma \cdot V_{Q',\pi}(j))| \\
&= \gamma \cdot \max_{j \in N(i,a)} |V_{Q,\pi}(j) - V_{Q',\pi}(j)| \\
&\leq \gamma \cdot \|V_{Q,\pi} - V_{Q',\pi}\| \\
&\leq \gamma \cdot \|Q - Q'\|.
\end{aligned}$$

This proves the second inequality. The proof of the first inequality is obtained from this proof by substituting B_γ , V_Q , and $V_{Q'}$ for B_γ^π , $V_{Q,\pi}$, and $V_{Q',\pi}$, respectively. ■

Appendix B

Proofs of the Results from Section 5

Proof of Theorem 13: The following four inequalities follow from Assumptions 1 and 4 in conjunction with the triangle inequality.

$$\|V - V_\gamma^*\| \leq \|V - T_\gamma V\| + \|T_\gamma V - V_\gamma^*\| \leq \|V - T_\gamma V\| + \gamma \cdot \|V - V_\gamma^*\|; \quad (\text{B.1})$$

$$\|V - T_\gamma V\| \leq \|V - V_\gamma^*\| + \|V_\gamma^* - T_\gamma V\| \leq \|V - V_\gamma^*\| + \gamma \cdot \|V_\gamma^* - V\|; \quad (\text{B.2})$$

$$\|V - V_\gamma^\pi\| \leq \|V - T_\gamma^\pi V\| + \|T_\gamma^\pi V - V_\gamma^\pi\| \leq \|V - T_\gamma^\pi V\| + \gamma \cdot \|V - V_\gamma^\pi\|; \quad (\text{B.3})$$

$$\|V - T_\gamma^\pi V\| \leq \|V - V_\gamma^\pi\| + \|V_\gamma^\pi - T_\gamma^\pi V\| \leq \|V - V_\gamma^\pi\| + \gamma \cdot \|V_\gamma^\pi - V\|. \quad (\text{B.4})$$

Inequalities (B.1), (B.2), (B.3), and (B.4) imply the first, second, third, and fourth inequality of this theorem, respectively. ■

Proof of Theorem 14: The first and third inequality have already been proven in the context of Theorem 13. From Assumptions 3, 1, and 6 we obtain

$$T_\gamma V - \|V - V_\gamma^*\| \leq T_\gamma V_\gamma^* - \|V - V_\gamma^*\| = V_\gamma^* - \|V - V_\gamma^*\| \leq V; \quad (\text{B.5})$$

$$T_\gamma^\pi V - \|V - V_\gamma^\pi\| \leq T_\gamma^\pi V_\gamma^\pi - \|V - V_\gamma^\pi\| = V_\gamma^\pi - \|V - V_\gamma^\pi\| \leq V. \quad (\text{B.6})$$

Assumptions 1 and 4 imply

$$\begin{aligned}
\|V_\gamma^* - T_\gamma V\| &= \|T_\gamma V_\gamma^* - T_\gamma V\| \leq \gamma \cdot \|V_\gamma^* - V\| \leq \|V - V_\gamma^*\|; \\
\|V_\gamma^\pi - T_\gamma^\pi V\| &= \|T_\gamma^\pi V_\gamma^\pi - T_\gamma^\pi V\| \leq \gamma \cdot \|V_\gamma^\pi - V\| \leq \|V - V_\gamma^\pi\|,
\end{aligned}$$

and therefore, in conjunction with Assumption 6,

$$V - T_\gamma V \leq V_\gamma^* - T_\gamma V \leq \|V_\gamma^* - T_\gamma V\| \leq \|V - V_\gamma^*\|; \quad (\text{B.7})$$

$$V - T_\gamma^\pi V \leq V_\gamma^\pi - T_\gamma^\pi V \leq \|V_\gamma^\pi - T_\gamma^\pi V\| \leq \|V - V_\gamma^\pi\|. \quad (\text{B.8})$$

Inequalities (B.5) and (B.7) imply $-\|V - V_\gamma^*\| \leq V - T_\gamma V \leq \|V - V_\gamma^*\|$ and thus the second inequality of the theorem. Inequalities (B.6) and (B.8) imply $-\|V - V_\gamma^\pi\| \leq V - T_\gamma^\pi V \leq \|V - V_\gamma^\pi\|$ and therefore the third inequality of the theorem. ■

Proof of Theorem 15: We prove the desired bound by applying, in turn, the triangle inequality, Assumptions 1, 5, and 4, Theorem 13, and finally Assumption 5 a second time:

$$\begin{aligned}
\|V_\gamma^\pi - V_\gamma^*\| &\leq \|V_\gamma^\pi - T_\gamma V\| + \|T_\gamma V - V_\gamma^*\| \\
&= \|T_\gamma^\pi V_\gamma^\pi - T_\gamma^\pi V\| + \|T_\gamma V - T_\gamma V_\gamma^*\| \\
&\leq \gamma \cdot \|V_\gamma^\pi - V\| + \gamma \cdot \|V - V_\gamma^*\| \\
&\leq \frac{\gamma}{1-\gamma} \cdot \|V - T_\gamma^\pi V\| + \frac{\gamma}{1-\gamma} \cdot \|V - T_\gamma V\| \\
&= \frac{\gamma}{1-\gamma} \cdot \|V - T_\gamma V\| + \frac{\gamma}{1-\gamma} \cdot \|V - T_\gamma V\| = \frac{2\gamma}{1-\gamma} \cdot d_B(V).
\end{aligned}$$

To see that this bound cannot be tightened any further, consider a deterministic MDP with two states, 1 and 2, and two actions, 1 and 2. Action 1 causes a state transition to state 1 no matter which state it is executed in. Similarly, action 2 always causes a state transition to state 2. All immediate costs are two, except for the execution of action 2 in state 2, which costs nothing. Now consider the value function V that is defined by $V(1) = V(2) = \frac{1}{1-\gamma}$. From Assumption 7 it follows that $T_\gamma V(1) - V(1) = 2 + \gamma V(1) - V(1) = 1$ and $T_\gamma V(2) - V(2) = 0 + \gamma V(2) - V(2) = -1$. Therefore $d_B(V) = \|T_\gamma V - V\| = 1$. The stationary policy π with $\pi(1) = 1$ and $\pi(2) = 2$ is greedy for V according to Assumptions 5 and 7. Assumptions 6 and 8 imply $V_\gamma^*(1) = 2$, $V_\gamma^\pi(1) = \frac{2}{1-\gamma}$, $V_\gamma^*(2) = 0$, and $V_\gamma^\pi(2) = 0$. It follows that $\|V_\gamma^\pi - V_\gamma^*\| = \frac{2\gamma}{1-\gamma} = \frac{2\gamma}{1-\gamma} \cdot d_B(V)$. ■

Proof of Theorem 16: We prove the desired bound by applying, in turn, the triangle inequality, Assumptions 1, 5, and 4, Theorem 13, Assumption 5 again, and finally Theorem 13 a second time:

$$\begin{aligned}
\|V_\gamma^\pi - V_\gamma^*\| &\leq \|V_\gamma^\pi - T_\gamma^\pi V\| + \|T_\gamma^\pi V - V_\gamma^*\| \\
&= \|T_\gamma^\pi V_\gamma^\pi - T_\gamma^\pi V\| + \|T_\gamma V - T_\gamma V_\gamma^*\| \\
&\leq \gamma \cdot \|V_\gamma^\pi - V\| + \gamma \cdot d_M(V) \\
&\leq \gamma \cdot \frac{\|V - T_\gamma^\pi V\|}{1-\gamma} + \gamma \cdot d_M(V) \\
&\leq \gamma \cdot \frac{(1+\gamma) \cdot \|V - V_\gamma^*\|}{1-\gamma} + \gamma \cdot d_M(V) \\
&= \gamma \cdot \frac{(1+\gamma) \cdot d_M(V)}{1-\gamma} + \gamma \cdot d_M(V)
\end{aligned}$$

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

To see that this bound cannot be tightened any further, consider again the Markov decision problem that we used in the proof of Theorem 15, but this time in conjunction with the value function V that is defined by $V(1) = V(2) = 1$. Assumptions 6 and 8 imply that $V_\gamma^*(1) = 2$ and $V_\gamma^*(2) = 0$. Since $V_\gamma^*(1) - V(1) = 2 - 1 = 1$ and $V_\gamma^*(2) - V(2) = 0 - 1 = -1$, it follows that $d_M(V) = \|V_\gamma^* - V\| = 1$. The stationary policy π with $\pi(1) = 1$ and $\pi(2) = 2$ is greedy for V according to Assumptions 5 and 7. Assumption 8 implies $V_\gamma^\pi(1) = \frac{2}{1-\gamma}$, $V_\gamma^\pi(2) = 0$, and, consequently, $\|V_\gamma^\pi - V_\gamma^*\| = \frac{2\gamma}{1-\gamma} = \frac{2\gamma}{1-\gamma} \cdot d_M(V)$. ■

Proof of Theorem 17: From Theorem 13 and Assumption 5 we obtain $\|V - V_\gamma^\pi\| \leq \|V - T_\gamma V\| \cdot \frac{1}{1-\gamma} = d_B(V) \cdot \frac{1}{1-\gamma}$. Therefore,

$$V \geq V_\gamma^\pi - d_B(V) \cdot \frac{1}{1-\gamma}. \quad (\text{B.9})$$

We can now prove the desired bound by applying, in turn, Assumptions 1, 3, and 5, Inequality (B.9), and finally Assumptions 3, 2, and 1.

$$\begin{aligned} V_\gamma^* &= T_\gamma V_\gamma^* \geq T_\gamma V = T_\gamma^\pi V \\ &\geq T_\gamma^\pi \left(V_\gamma^\pi - \frac{d_B(V)}{1-\gamma} \right) = T_\gamma^\pi V_\gamma^\pi - \frac{\gamma}{1-\gamma} \cdot d_B(V) \\ &= V_\gamma^\pi - \frac{\gamma}{1-\gamma} \cdot d_B(V). \end{aligned}$$

Together with Assumption 6 we obtain $0 \leq V_\gamma^\pi - V_\gamma^* \leq \frac{\gamma}{1-\gamma} \cdot d_B(V)$ and, therefore, $\|V_\gamma^\pi - V_\gamma^*\| \leq \frac{\gamma}{1-\gamma} \cdot d_B(V)$.

To see that this bound cannot be tightened any further, consider again the MDP that we used in the proof of Theorem 15, but this time in conjunction with the value function V that is defined by $V(1) = V(2) = 0$. Assumptions 6 and 8 imply that $V_\gamma^*(1) = 2$ and $V_\gamma^*(2) = 0$ and, therefore, $V_\gamma^*(1) = 2 \geq V(1)$ and $V_\gamma^*(2) = 0 \geq V(2)$. From Assumption 7 it follows that

$$T_\gamma V(1) - V(1) = 2 - 0 = 2; \quad (\text{B.10})$$

$$T_\gamma V(2) - V(2) = 0 - 0 = 0. \quad (\text{B.11})$$

Therefore $d_B(V) = \|T_\gamma V - V\| = 2$. The stationary policy π with $\pi(1) = 1$ and $\pi(2) = 2$ is greedy for V according to Assumptions 5 and 7, and Assumption 8 implies $V_\gamma^\pi(1) = \frac{2}{1-\gamma}$ and $V_\gamma^\pi(2) = 0$. We can conclude that $\|V_\gamma^\pi - V_\gamma^*\| = \frac{2\gamma}{1-\gamma} = \frac{\gamma}{1-\gamma} \cdot d_B(V)$. ■

Proof of Corollary 1: The desired bound is an immediate consequence of Theorem 17 in conjunction with the second inequality of Theorem 14. The example used to prove

Theorem 17 proves that the bound is tight, because it satisfies $d_B(V) = d_M(V)$. ■

LEMMA 1 *Let V be a value function that satisfies $V \leq T_\gamma V$. Then $V \leq V_\gamma^*$.*

Proof: Consider mappings $T_{\gamma,n}$ from $\mathcal{R}^S$ to $\mathcal{R}^S$ that are inductively defined as follows:

$$\forall n \in \mathbb{N} \forall V \in \mathcal{R}^S : T_{\gamma,0}V = V, \quad T_{\gamma,n+1}V = T_\gamma(T_{\gamma,n}V).$$

It follows by induction from Assumption 3 that $V \leq T_{\gamma,n}V$ for all n if $V \leq T_\gamma V$. Similarly, it follows by induction from Assumption 1 that $T_{\gamma,n}V_\gamma^* = V_\gamma^*$ and from Assumption 4 that $\|T_{\gamma,n}V_\gamma^* - T_{\gamma,n}V\| \leq \gamma^n \|V_\gamma^* - V\|$, for all n . Put together, we can conclude that $\|V_\gamma^* - T_{\gamma,n}V\| \leq \gamma^n \|V_\gamma^* - V\|$ for all n which further implies $\lim_{n \rightarrow \infty} T_{\gamma,n}V = V_\gamma^*$ and, finally, $V \leq V_\gamma^*$. ■

Proof of Corollary 2: The desired bound is an immediate consequence of Lemma 1 and Theorem 17. The example used to prove Theorem 17 proves that the bound is tight, because it satisfies $V \leq T_\gamma V$ according to Equations (B.10) and (B.11). ■

Proof of Corollary 3: The desired bound follows immediately from Lemma 1 and Corollary 1. The example used to prove Theorem 17 proves that the bound is tight, because it satisfies $V \leq T_\gamma V$ according to Equations (B.10) and (B.11) and, furthermore, $d_B(V) = d_M(V)$. ■

Proof of Theorem 18: The following four inequalities follow from Assumptions 10 and 13 in conjunction with the triangle inequality.

$$\|Q - Q_\gamma^*\| \leq \|Q - B_\gamma Q\| + \|B_\gamma Q - Q_\gamma^*\| \leq \|Q - B_\gamma Q\| + \gamma \cdot \|Q - Q_\gamma^*\|; \quad (\text{B.12})$$

$$\|Q - B_\gamma Q\| \leq \|Q - Q_\gamma^*\| + \|Q_\gamma^* - B_\gamma Q\| \leq \|Q - Q_\gamma^*\| + \gamma \cdot \|Q_\gamma^* - Q\|; \quad (\text{B.13})$$

$$\|Q - Q_\gamma^\pi\| \leq \|Q - B_\gamma^\pi Q\| + \|B_\gamma^\pi Q - Q_\gamma^\pi\| \leq \|Q - B_\gamma^\pi Q\| + \gamma \cdot \|Q - Q_\gamma^\pi\|; \quad (\text{B.14})$$

$$\|Q - B_\gamma^\pi V\| \leq \|Q - Q_\gamma^\pi\| + \|Q_\gamma^\pi - B_\gamma^\pi Q\| \leq \|Q - Q_\gamma^\pi\| + \gamma \cdot \|Q_\gamma^\pi - Q\|. \quad (\text{B.15})$$

Inequalities (B.12), (B.13), (B.14), and (B.15) imply the first, second, third, and fourth inequality of this theorem, respectively. ■

Proof of Theorem 19: The first and third inequality have already been proven in the context of Theorem 18. From Assumptions 12 and 10, we obtain

$$B_\gamma Q - \|Q - Q_\gamma^*\| \leq B_\gamma Q_\gamma^* - \|Q - Q_\gamma^*\| = Q_\gamma^* - \|Q - Q_\gamma^*\| \leq Q; \quad (\text{B.16})$$

$$B_\gamma^\pi Q - \|Q - Q_\gamma^\pi\| \leq B_\gamma^\pi Q_\gamma^\pi - \|Q - Q_\gamma^\pi\| = Q_\gamma^\pi - \|Q - Q_\gamma^\pi\| \leq Q. \quad (\text{B.17})$$

Assumptions 10 and 13 imply

$$\begin{aligned}\|Q_\gamma^* - B_\gamma Q\| &= \|B_\gamma Q_\gamma^* - B_\gamma Q\| \leq \gamma \cdot \|Q_\gamma^* - Q\| \leq \|Q - Q_\gamma^*\|, \\ \|Q_\gamma^\pi - B_\gamma^\pi Q\| &= \|B_\gamma^\pi Q_\gamma^\pi - B_\gamma^\pi Q\| \leq \gamma \cdot \|Q_\gamma^\pi - Q\| \leq \|Q - Q_\gamma^\pi\|,\end{aligned}$$

and therefore, in conjunction with Assumption 15

$$Q - B_\gamma Q \leq Q_\gamma^* - B_\gamma Q \leq \|Q_\gamma^* - B_\gamma Q\| \leq \|Q - Q_\gamma^*\|; \quad (\text{B.18})$$

$$Q - B_\gamma^\pi Q \leq Q_\gamma^* - B_\gamma^\pi Q \leq Q_\gamma^\pi - B_\gamma^\pi Q \leq \|Q_\gamma^\pi - B_\gamma^\pi Q\| \leq \|Q - Q_\gamma^\pi\|. \quad (\text{B.19})$$

Inequalities (B.16) and (B.18) imply $\|Q - Q_\gamma^*\| \leq \|Q - B_\gamma Q\|$ and thus the second inequality of the theorem. Inequalities (B.17) and (B.19) imply $\|Q - Q_\gamma^\pi\| \leq \|Q - B_\gamma^\pi Q\|$ and therefore the third inequality of the theorem. ■

Proof of Theorem 20: We prove the desired bound by applying, in turn, the triangle inequality, Assumption 9, Theorems 3, 6 and 18, and finally Assumption 14:

$$\begin{aligned}\|V_\gamma^\pi - V_\gamma^*\| &\leq \|V_\gamma^\pi - V_Q\| + \|V_Q - V_\gamma^*\| \\ &= \|V_{Q_\gamma^\pi, \pi} - V_{Q, \pi}\| + \|V_Q - V_{Q_\gamma^*}\| \\ &\leq \|Q_\gamma^\pi - Q\| + \|Q - Q_\gamma^*\| \\ &\leq \frac{\|Q - B_\gamma^\pi Q\|}{1 - \gamma} + \frac{\|Q - B_\gamma Q\|}{1 - \gamma} \\ &= \frac{2\|Q - B_\gamma Q\|}{1 - \gamma} = \frac{2}{1 - \gamma} \cdot d_B(Q).\end{aligned}$$

To see that this bound cannot be tightened any further, consider a deterministic MDP with a single state 1 and two actions, 1 and 2, that both cause self transitions and have immediate costs 2 and 0, respectively. Now consider the Q-function Q that is defined by $Q(1, 1) = Q(1, 2) = 1/(1 - \gamma)$. From Assumption 16 it follows that $B_\gamma Q(1, 1) - Q(1, 1) = 2 + \gamma \cdot V_Q(1) - Q(1, 1) = 2 + \frac{\gamma}{1 - \gamma} - \frac{1}{1 - \gamma} = 1$ and $B_\gamma Q(1, 2) - Q(1, 2) = 0 + \gamma \cdot V_Q(1) - Q(1, 2) = \frac{\gamma}{1 - \gamma} - \frac{1}{1 - \gamma} = -1$. Therefore $d_B(Q) = \|B_\gamma Q - Q\| = 1$. The stationary policy π with $\pi(1) = 1$ is greedy for Q , and Assumptions 6 and 8 imply $V_\gamma^\pi(1) = 2/(1 - \gamma)$ and $V_\gamma^*(1) = 0$. It follows that $\|V_\gamma^\pi - V_\gamma^*\| = 2/(1 - \gamma) = 2d_B(Q)/(1 - \gamma)$. ■

Proof of Theorem 21: We prove the desired bound by applying, in turn, the triangle inequality, Assumption 9, Theorems 3, 6 and 18, Assumption 14, and finally Theorem 18 a second time:

$$\begin{aligned}\|V_\gamma^\pi - V_\gamma^*\| &\leq \|V_\gamma^\pi - V_{Q, \pi}\| + \|V_Q - V_\gamma^*\| \\ &= \|V_{Q_\gamma^\pi, \pi} - V_Q\| + \|V_Q - V_{Q_\gamma^*}\|\end{aligned}$$

$$\begin{aligned}
&\leq \|Q_\gamma^\pi - Q\| + \|Q - Q_\gamma^*\| \\
&\leq \frac{\|Q - B_\gamma^\pi Q\|}{1 - \gamma} + \|Q - Q_\gamma^*\| \\
&= \frac{\|Q - B_\gamma^\pi Q\|}{1 - \gamma} + d_M(Q) \\
&\leq \frac{(1 + \gamma) \cdot d_M(Q)}{1 - \gamma} + d_M(Q) \\
&= \frac{2}{1 - \gamma} \cdot d_M(Q).
\end{aligned}$$

To see that this bound cannot be tightened any further, consider again the MDP that we used in the proof of Theorem 20, but this time in conjunction with the Q-function Q that is defined by $Q(1, 1) = Q(1, 2) = 1$. The stationary policy π defined by $\pi(1) = 1$ is greedy for Q and Assumptions 6 and 8 imply that $V_\gamma^\pi(1) = 2/(1 - \gamma)$ and $V_\gamma^*(1) = 0$. Assumptions 9, 10, and 16 imply $Q_\gamma^*(i, a) = c_i^a + \gamma \cdot V_{Q_\gamma^*}(j_i^a) = c_i^a + \gamma \cdot V_\gamma^*(j_i^a)$ for all deterministic MDPs, states i , and actions $a \in A(i)$. Hence, $Q_\gamma^*(1, 1) = 2$ and $Q_\gamma^*(1, 2) = 0$ for the example. It follows that $d_M(Q) = \|Q_\gamma^* - Q\| = 1$ and, finally, $\|V_\gamma^\pi - V_\gamma^*\| = 2/(1 - \gamma) = 2d_M(Q)/(1 - \gamma)$. ■

Proof of Theorem 22: Let i be an arbitrary state. We prove the desired bound by applying, in turn, Assumptions 6 and 9, Theorems 2, 3, 6, and 18, and finally Assumption 14.

$$\begin{aligned}
\|V_\gamma^\pi - V_\gamma^*\| &\leq V_\gamma^\pi(i) - V_\gamma^*(i) \\
&= V_{Q_\gamma^\pi, \pi}(i) - V_{Q_\gamma^*}(i) \leq V_{Q_\gamma^\pi, \pi}(i) - V_Q(i) \\
&= V_{Q_\gamma^\pi, \pi}(i) - V_{Q, \pi}(i) \leq \|V_{Q_\gamma^\pi, \pi} - V_{Q, \pi}\| \\
&\leq \|Q_\gamma^\pi - Q\| \leq \frac{\|Q - B_\gamma^\pi Q\|}{1 - \gamma} \\
&= \frac{\|Q - B_\gamma Q\|}{1 - \gamma} = \frac{d_B(Q)}{1 - \gamma}.
\end{aligned}$$

To see that this bound cannot be tightened any further, consider again the MDP that we used in the proof of Theorems 20 and 21, but this time in conjunction with the Q-function Q that is defined by $Q(1, 1) = Q(1, 2) = 0$. In the proof of Theorem 21, we showed that $V_\gamma^*(1) = 0$, $Q_\gamma^*(1, 1) = 2$, and $Q_\gamma^*(1, 0) = 0$. Therefore, $Q \leq Q_\gamma^*$. Assumption 16 implies that $B_\gamma Q(1, 1) - Q(1, 1) = 2 + \gamma \cdot V_Q(1) - Q(1, 1) = 2$, $B_\gamma Q(1, 2) - Q(1, 2) = 0 + \gamma \cdot V_Q(1) - Q(1, 2) = 0$, and therefore $d_B(Q) = \|B_\gamma Q - Q\| = 2$. The stationary policy π with $\pi(1) = 1$ is greedy for Q . Assumptions 6 and 8 imply $V_\gamma^\pi(1) = 2/(1 - \gamma)$, and, therefore, $\|V_\gamma^\pi - V_\gamma^*\| = 2/(1 - \gamma) = d_B(Q)/(1 - \gamma)$. ■

Proof of Corollary 4: The desired bound is an immediate consequence of Theorems 22 and 19. The example used to prove Theorem 22 proves that the bound is tight, because it satisfies $Q_\gamma^*(1, 1) = 2$, $Q_\gamma^*(1, 2) = 0$, and therefore $d_B(Q) = d_M(Q)$. ■

LEMMA 2 *Let Q be a Q -function that satisfies $Q \leq B_\gamma Q$. Then $Q \leq Q_\gamma^*$.*

Proof: Consider mappings $B_{\gamma,n}$ from $\mathcal{R}^M$ to $\mathcal{R}^M$ that are inductively defined as follows:

$$\forall n \in \mathbb{N} \forall Q \in \mathcal{R}^M : B_{\gamma,0}Q = Q, \quad B_{\gamma,n+1}Q = B_\gamma(B_{\gamma,n}Q).$$

It follows by induction from Assumption 12 that $Q \leq B_{\gamma,n}Q$ for all n if $Q \leq B_\gamma Q$. Similarly, it follows by induction from Assumption 10 that $B_{\gamma,n}Q_\gamma^* = Q_\gamma^*$ and from Assumption 13 that $\|B_{\gamma,n}Q_\gamma^* - B_{\gamma,n}Q\| \leq \gamma^n \|Q_\gamma^* - Q\|$, for all n . Put together, we can conclude that $\|Q_\gamma^* - B_{\gamma,n}Q\| \leq \gamma^n \|Q_\gamma^* - Q\|$ for all n which further implies $\lim_{n \rightarrow \infty} B_{\gamma,n}Q = Q_\gamma^*$ and, finally, $Q \leq Q_\gamma^*$. ■

Proof of corollary 5: Lemma 2 implies $Q \leq Q_\gamma^*$. Thus, Theorem 22 can be applied and gives the desired result. ■

Proof of Corollary 6: Lemma 2 implies $Q \leq Q_\gamma^*$. Thus, Corollary 4 can be applied and gives the desired result. ■

Proof of Theorem 23: Let Q be an arbitrary Q -function. We will show that for every state i there exists a real value $\epsilon(i) > 0$ with the following property: If $\max_{a \in A(i)} |Q(i, a) - Q_\gamma^*(i, a)| < \epsilon(i)$ then every action that is greedy for Q and i is also greedy for Q_γ^* and i . If such values $\epsilon(i)$ exist, then $\epsilon := \min_{i \in S} \epsilon(i)$ gives the desired result.

Let i be any state. If every action in $A(i)$ is greedy for Q_γ^* and i , then $\epsilon(i)$ can be set to an arbitrary non-zero value, for example one. Otherwise, set

$$\epsilon(i) := \frac{\min \left\{ r \in \mathcal{R} : \exists a \in A(i) : r = Q_\gamma^*(i, a) \neq V_{Q_\gamma^*}(i) \right\}}{2}.$$

The desired result follows using Assumption 17. ■

Proof of Corollary 7: The corollary follows immediately from Theorems 23 and 18. ■

Proof of Theorem 24: From Assumption 17 we conclude that the Q -function $Q_\gamma^* + 2r$ has the desired property. ■

Proof of Corollary 8: The corollary follows immediately from Theorems 24 and 18. ■

Notes

1. In the dynamic programming literature, it is often assumed that the immediate cost of a state transition depends only on the starting state and the executed action, but not on the successor state reached. This is the assumption that we make in this section. We could easily extend the theory to the more general case, but are afraid that the resulting dynamic programming operators might look unfamiliar to the reader.
2. This assumption is made for convenience. It follows already from Assumptions 1 and 7.

References

- Baird, L., (1995). "Residual Algorithms: Reinforcement Learning with Function Approximation," *Machine Learning, Proceedings of the 12th International Conference*, Morgan Kaufmann Publishers, Inc., San Francisco, CA.
- Barto, A. G., Sutton, R. S., & Anderson, C. W., (1983). "Neuronlike adaptive elements that can solve difficult learning control problems," *IEEE Transactions on Systems, Man, and Cybernetics* 13, pp. 834-846.
- Bertsekas, D.P., (1987). *Dynamic Programming: Deterministic and Stochastic Models*, Englewood Cliffs, NJ: Prentice Hall.
- Boyan, J. & Moore, A., (1995). "Generalization in Reinforcement Learning: Safely Approximating the Value Function," G. Tesauro, D. S. Touretzky and T. K. Lean, editors, *Advances in Neural Information Processing Systems* 7, MIT Press, Cambridge, MA.
- Crites, R. H. & Barto, A. G., (1995, unpublished). "Improving Elevator Performance Using Reinforcement Learning."
- Gordon, G., (1995). "Stable function approximation in Dynamic Programming," *Machine Learning, Proceedings of the 12th International Conference*, Morgan Kaufmann Publishers, Inc., San Francisco, CA.
- Heger, M. (1994a), "Consideration of Risk in Reinforcement Learning," *Machine Learning, Proceedings of the 11th International Conference*, pp. 105-111, Morgan Kaufmann Publishers, Inc., San Francisco, CA.
- Heger, M. (1994b), "Risk and Reinforcement Learning: Concepts and Dynamic Programming," *Technical report No. 8/94*, Zentrum für Kognitionswissenschaften, Universität Bremen, Fachbereich 3 Informatik, Germany.
- Heger, M. (1995), "Risk and Reinforcement Learning: Algorithms," *Technical report (forthcoming)*, Zentrum für Kognitionswissenschaften, Universität Bremen, Fachbereich 3 Informatik, Germany.
- Lin, L.-J., (1993). "Scaling Up Reinforcement Learning for Robot Control," *Machine Learning, Proceedings of the 10th International Conference*, pp. 182-189, Morgan Kaufmann Publishers, Inc., San Francisco, CA.
- Littman, M. L., (1994). "Markov games as a framework for multi-agent reinforcement learning," *Machine Learning, Proceedings of the 11th International Conference*, pp. 157-111, Morgan Kaufmann Publishers, Inc., San Francisco, CA.
- McDonald, M. A. F. & Hingson P., (1994). "Approximate Discounted Dynamic Programming Is Unreliable," *Technical report 94/6*, University of Western Australia.
- Moore, A.W. & Atkeson, C. G., (to appear). "The Parti-game Algorithm for Variable Resolution Reinforcement Learning in Multidimensional State-spaces," *Machine Learning*.
- Ross, S. M., (1970). *Applied Probability Models with Optimization Applications*, San Francisco, California: Holden Day.
- Schwartz, A., (1993). "A reinforcement learning method for maximizing undiscounted rewards," *Machine Learning, Proceedings of the 10th International Conference*, pp. 298-305, Morgan Kaufmann Publishers, Inc., San Francisco, CA.
- Singh, P. S. & Yee, R. C., (1994). "An Upper Bound on the Loss from Approximate Optimal-Value Functions," *Machine Learning*, No 16, pp. 227-233, Kluwer Academic Publishers, Boston.

- Sutton, R. S., (unpublished). "Generalization in Reinforcement Learning: Successful Examples Using Sparse Coarse Coding."
- Taha, H. A., (1987). *Operations Research: An Introduction, Fourth Edition*, Macmillan Publishing Company, New York.
- Tesauro, G. J., (1992). "Practical Issues in Temporal Difference Learning," *Machine Learning*, 8(3/4), pp. 257-277.
- Thrun, S. B., (1992) "Efficient Exploration in Reinforcement Learning," *Technical report, CMU-CS-92-102*, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA.
- Thrun, S. B. & Schwartz, A., (1993). "Issues in Using Function Approximation for Reinforcement Learning," *Proceedings of the Fourth Connectionist Models Summer School*, Lawrence Erlbaum Publishers, Hillsdale, NJ.
- Tsitsiklis, J. N., (1994). "Asynchronous Stochastic Approximation and Q-Learning," *Machine Learning*, No 16, pp. 185-202, Kluwer Academic Publishers, Boston.
- Watkins, C. J. C. H., (1989). *Learning from Delayed Rewards*, PhD thesis, Cambridge University, England.
- Watkins, C. J. C. H. & Dayan, P., (1992). "Q-Learning," *Machine Learning*, No 8, pp. 279-292.
- Williams, R.J. & Baird, III, L.C., (1993). "Tight Performance Bounds on Greedy Policies Based on Imperfect Value Functions," *Technical Report NU-CCS-93-14*, Northeastern University.
- Zhang, W. & Dietterich, T. G., (1995). "A Reinforcement Learning Approach to Job-shop Scheduling," *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Morgan Kaufmann Publishers, Inc., San Francisco, CA.

Received November 8, 1994

Accepted February 24, 1995

Final Manuscript October 6, 1995