Consideration of Risk in Reinforcement Learning

Matthias Heger
Zentrum für Kognitionswissenschaften
Universität Bremen, FB3 Informatik
Postfach 330 440
D-28334 Bremen, Germany
heger@informatik.uni-bremen.de

Abstract

Most Reinforcement Learning (RL) work supposes policies for sequential decision tasks to be optimal that minimize the expected total discounted cost (e.g. Q-Learning; AHC architecture). On the other hand, it is well known that it is not always reliable and can be treacherous to use the expected value as a decision criterion. A lot of alternative decision criteria have been suggested in decision theory to get a more sophisticated consideration of risk but most RL researchers have not concerned themselves with this subject until now. The purpose of this paper is to draw the reader's attention to the problems of the expected value criterion in Markov decision processes and to give Dynamic Programming algorithms for an alternative criterion, namely the minimax criterion. A counterpart to Watkins' Q-Learning with regard to the minimax criterion is presented. The new algorithm, called $\hat{Q}$-learning, finds policies that minimize the worst-case total discounted cost.

1 INTRODUCTION

In the paradigm of *Reinforcement Learning* (RL), an agent interacts with an environment (world) and, simultaneously, receives reinforcement signals which are punishments or costs caused by single decisions and/or state transitions[1]. The learning task is to find a favorable policy, i.e., a rule that tells the agent which action to choose in a given situation. The most frequently used model of the interaction between the agent and the world is a special kind of stochastic process, namely the *Markov decision process* (MDP).

Three sets are elementary in MDP: The set S of the environment's states *from the agent's point of view*, the set A of the agent's actions and the set C of the reinforcement signals which, in the following, will be also called immediate costs. In this paper we confine ourself to finite sets S and A represented by subsets of N. The elements of $C \subset \mathrm{R}$ are assumed to be countable, bounded and nonnegative. In general, the agent must not or is not able to select an arbitrary action in every state. The nonempty set $A(i) \subseteq A$ denotes the set of admissible actions in state i.

The agent's interaction with the world is divided into a sequence of so-called stages or *episodes*. In the following, a paradigm of an episode is given: First, the agent observes a *starting state* $i \in S$ and has to choose and perform an *action* $a \in A(i)$. Then, a state transition occurs from state i into a *successor state* $j \in S$. Finally, the agent receives a scalar *reinforcement signal* $r \in C$. The *immediate cost* r represents the effort of the state transition from i to j under action a.

It is essential in MDP that at any time t, the probability that the successor state of episode t equals a certain state j depends only on the starting state i and the action a of that episode but not additionally on the time and past episodes. $P_S(i, a, j)$ represents the probability that the successor state of an episode is j for a given starting state i and action a.

Similarly, at any time t, the probability that the immediate cost of episode t equals a certain number $r \in C$ depends only on the starting state, the action and the successor state of that episode but neither additionally on the time nor on 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.

In MDP, the starting state of episode t equals the successor state of episode $t-1$ for all $t > 0$. The starting state of episode zero may be given by a probability distribution. $P_0(i)$ denotes the probabilty that i is the starting state of the episode with the time index $t = 0$. For technical reasons we suppose $P_0(i) > 0$ for all $i \in S$.

The agent's behavior is specified by a *policy* which is a rule that yields an action in any given situation. In a comprehensive model of policies, the agent may consider at any time the current state, the time and the past states and actions

[1] In literature, reinforcement signals often also represent rewards, but it is well known that the reward representation can easily be transformed into the cost representation and vice versa.

in order to select an action. Additionally, the selection of the action may be done probabilistically. Action selection by the so called *stationary policies* only require the agent to consider the current state. These policies identify with mere mappings from states into actions but prove to be very powerful.

States and immediate costs are random variables in MDP because of the probabilistic nature of the immediate costs and state transitions. These random variables essentially depend on the policy applied by the agent. Therefore, the following notation is used: I_t^π denotes the starting state and the C_t^π the immediate cost where π and t describe the agent's policy and the time index of the episode, respectively.

2 PROBLEMS OF THE EXPECTED RETURN

In this section we introduce the commonly used measure of performance for policies and emphasize some drawbacks that are known from decision theory.

2.1 THE EXPECTED RETURN

For the remainder of the text assume $\gamma \in [0, 1)$. We concentrate on

$$R_{\gamma,0}^\pi := 0, \quad R_{\gamma,t}^\pi := \sum_{\tau=0}^{t-1} \gamma^\tau C_\tau^\pi \quad \text{and} \quad R_\gamma^\pi := \sum_{\tau=0}^{\infty} \gamma^\tau C_\tau^\pi$$

where π is a policy and $t \in \mathrm{N}^+$.

R_γ^π is called the *(discounted) return (of π)* or the *total (discounted) cost*. By the return all immediate costs are taken into account and because of the *discount factor* γ, costs are weighted the less the more distant they lie in the future. The total discounted cost is to be found often in RL because a lot of problems as, e.g., *goal tasks* or *time-until-success tasks* and *time-until-failure tasks* (Sutton 1984) can be represented by the task of finding a policy which minimizes the total cost. $R_{\gamma,t}^\pi$ is the *t-step (discounted) return (of π)* and is important because it may serve as an approximation for R_γ^π.

The return by itself is not an ordered measure of performance for policies and it is not trivial to derive one from the return. The problem arises from the fact that the return R_γ^π is not a real number but a *random variable*. In deterministic domains, random variables identify with real numbers and, undoubtedly, in such domains the agent should find a policy that minimizes the total cost. But what policy is to be preferred in probabilistic domains? The usual answer for this problem in RL is to concentrate on the expected value of the return as it is also done in Q-learning (Watkins, 1989) and the AHC architecture (Barto, Sutton & Anderson, 1983; Sutton, 1991). To be more precisely, most RL researchers measure the performance of a policy π by $\mathrm{E}\left(R_\gamma^\pi \mid I_0^\pi = i\right)$, i.e., the *expected return* relating to a given starting state i of episode 0. π is usually called optimal if it minimizes $\mathrm{E}\left(R_\gamma^\pi \mid I_0^\pi = i\right)$ for each $i \in S$. In operations research and decision theory (e.g. Taha, 1987), however, it is well known that it is not always reliable and can be treacherous to use the expected value as a decision criterion.

2.2 PROBLEMS OF THE DECISION CRITERION OF EXPECTATION

We give three simple examples from decision theory (Bradley, 1976) that show different problems of the criterion of expectation. The first problem is revealed by the celebrated *St. Petersburg Paradox*. A man offers you the privilege to play the following game for a stake of k dollars: A fair coin is to be tossed repeatedly until a head comes up. If it comes up head on the first toss, your opponent pays you two dollars. If a head first comes up on the second toss you get four dollars. And, in general, if the first head comes up on the nth toss, you receive 2^n dollars. The probability that the first head comes up on the nth toss is $1/2^n$. The expected value of the game's payoff X is $\mathrm{E}(X) = \sum_{n=1}^{\infty} \frac{1}{2^n} 2^n - k = \infty - k = \infty$.

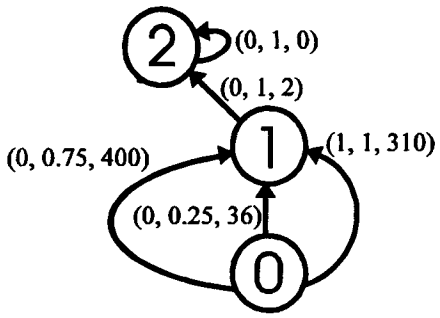
If you decide not to play the game, the expected payoff is of course zero. Therefore, by the expected value as a decision criterion, you have to decide to play the game for *any* finite stake. But no one would put up an arbitrary high stake to play a game in which recouping one's stake has infinitesimal small probabilities. The amount of the possible gain is appreciable enough but the probability of obtaining it is too small.

The second example is a lottery in which there are 10,001 tickets, of which 10,000 are winners and only one is a loser. It costs \$100 to buy a ticket; if you win you get \$100.01 and if you lose you get nothing. The expected value for the payoff X of the game is $\mathrm{E}(X) = \frac{10{,}000}{10{,}001}(\$100.01) + \frac{1}{10{,}001}(\$0) - \$100 = \$0$.

Hence, by the criterion of expectation it is equivalent to decide to play or not to play the game. On the other hand, most people would undoubtedly decide not to play the game because the possible gain of one cent is too small to risk the loss of the stake of \$100. This decision is hardly influenced by the fact that the probability for the loss is very small.

Finally, consider the plight of a businessman who has \$1,000 cash and no credit and who must raise an additional \$100,000 immediately or lose his business. A gambler offers him the following gamble: the businessman makes a single throw with a pair of fair dice; if the dice come up 2 or 12, the gambler pays the businessman \$10,000, but if any other number comes up, the businessman pays the gambler \$1,000. The expected value for the payoff X of the game is $\mathrm{E}(X) = \$10,000\left(\frac{1}{18}\right) + (-\$1,000)\left(\frac{17}{18}\right) = \$388.89$.

By the criterion of expectation, the businessman has to decide not to play the game. Yet it is definitely to his advantage to take the gamble. Subjective values are not considered by the criterion of expectation. In this example, the problem of



$x \in S$	$\pi(x) \in A(x)$	$\mu(x) \in A(x)$
0	0	1
1	0	0
2	0	0

Figure 1: In this MDP, denoted by a state transition graph, there are three states (circles). A state transition from a starting state i into a successor state j is represented by a directed edge labeled by a triple. The first number a in the triple is an admissible action for state i. The second number denotes the probability that the state transition will occur if action a is selected in the corresponding starting state i. Finally, the third number represents the immediate cost for the transition. The table denotes the only two possible stationary policies π and μ.

subjective values could be solved by a simple redefinition of the decision's outcome. In section 2.3, we will see that in MDP this problem is more subtle.

We give an extension of the summary from (Bradley, 1976): The expected value as a criterion for action in choice behavior

(a) is based upon long-run considerations where the decision process is repeated a sufficiently large number of times under same conditions. It is not necessarily a valid criterion in the short run or one-shot case, especially when either the possible consequences or their probabilities have extreme values.

(b) assumes the subjective values of the possible outcomes are proportional to their objective values, which is not necessarily the case, especially when the values involved are large.

2.3 THE CRITERION OF EXPECTATION IN MDP

The decision problem, we are confronted with in MDP, is the question what policy is to be used by the agent if the current state i of the world is given. In our framework, the outcome of a decision for a policy is its return.

Since all immediate costs of the future are involved by the return, it needs an infinite amount of time until the outcome of the decision is present. Therefore the agent has no time to repeat the decision process, and there is by definition no possibility to satisfy the long-run condition as mentioned in (a). In practice, however, this does not imply that long-run considerations are generally out of the question because the return may be approximated by the t-step return where $t \in \mathrm{N}$ is assumed to be sufficiently large.

The more important issue whether long-run considerations are adequate to measure the performance of a policy with respect to a starting state i exists in the condition that the agent shall and will visit i sufficiently often. This condition essentially depends on the domain and the agent's task. In the MDP of figure 1, e.g., there is no admissible sequence of actions that will eventually bring the world into state 0 more than once. In this domain, long-run considerations with regard to state 0 are adequate only then if one assumes events not involved in the MDP that may transfer the world into state 0 sufficiently often.

In the following, the significance of (b) is to be considered, i.e., the problem of subjective values of the decision's outcome. Assume a so-called utility function $U \in \mathrm{R}^{\mathrm{R}}$ that maps the value of the decision's outcome, i.e. the return, into a subjective value. In general, we have

$$U\left(R_\gamma^\pi\right) \;=\; U\left(\sum_{\tau=0}^{\infty}\gamma^\tau C_\tau^\pi\right) \;\neq\; \sum_{\tau=0}^{\infty}\gamma^\tau U\left(C_\tau^\pi\right).$$

Therefore, the problem of maximizing the subjective value (utility) of the return is generally not equivalent to the task to maximize the discounted sum of utilities of the immediate costs. Hence, the problem of subjective values is not merely to be solved by an easy redefinition of the underlying model of the immediate costs in the MDP. In (Koenig & Simmons, 1993) this problem is discussed for planning tasks.

3 DECISION CRITERIA

Decision theory offers several alternatives to the decision criterion of expectation in order to have a more sophisticated consideration of risk. Two categories of decision-making situations are to be considered in nondeterministic domains: *Decision under risk* and *decision under uncertainty*. In the first category all probability distributions that specify the domain are known, whereas in the second category no probability density function is available. Since we concern ourself with the definition of an *optimal* (decision for a) policy, the definition should be based on all information that specify the domain of the MDP. Therefore, we concentrate on decision under risk.

To give a very short list of some criteria of decision under risk, we use, as an example, our decision problem of what

policy to be used by the agent if the current state i of the world is given. For further examples of possible criteria and a discussion of the following criteria see e.g. (Taha, 1987).

- As mentioned above, by the *expected value criterion* a policy π is to be selected that minimizes $\mathrm{E}\left(R_\gamma^\pi \mid I_0^\pi = i\right)$.
- The *expected utility criterion* comes from utility theory (von Neumann & Morgenstern, 1947). An utility function $U \in \mathrm{R}^{\mathrm{R}}$ is to be assumed that maps the value of the return into a subjective value. A policy π is to be selected that maximizes $\mathrm{E}\left(U\left(R_\gamma^\pi\right) \mid I_0^\pi = i\right)$. Problem (a) and (b) of the criterion of expectation can be considered by the expected utility criterion. Unfortunately, although guidelines for establishing utility functions have been developped, utility is a rather subtle concept that cannot be quantified easily (Taha, 1987).
- By the *expected value-variance* criterion a policy π is to be chosen that minimizes $\mathrm{E}\left(R_\gamma^\pi \mid I_0^\pi = i\right) - K \cdot \mathrm{var}\left(R_\gamma^\pi \mid I_0^\pi = i\right)$ where K is a prespecified constant which is sometimes referred to as *risk aversion factor*. Because in our framework reinforcement signals are costs, K is assumed to be negative.
- Finally, by the *α-value criterion* a constant $\alpha \in [0, 1)$ is to be prespecified and the agent's policy π shall minimize the *α-value*

$$\begin{aligned} \hat{m}_\alpha &:= \hat{m}_\alpha\left(R_\gamma^\pi \mid I_0^\pi = i\right) \\ &:= \sup\{r \in \mathrm{R} : P\left(R_\gamma^\pi > r \mid I_0^\pi = i\right) > \alpha\} \end{aligned}$$

  where $P\left(R_\gamma^\pi > r \mid I_0^\pi = i\right)$ denotes the probability that the return is greater than r if the starting state i of episode 0 is given. In (Heger, 1994) it is shown that $P\left(R_\gamma^\pi \leq \hat{m}_\alpha \mid I_0^\pi = i\right) \geq 1 - \alpha$. Therefore the agent which starts in state i and uses policy π has the security that, at least with probability 1-α, the total cost will not exceed $\hat{m}_\alpha$. On the other hand, the agent has to be aware of the risk that the total cost will exceed any number less than $\hat{m}_\alpha$ with a probability greater than α. See (Heger, 1994) for an extended discussion of this criterion.

For the remainder of the text we restrict our attention to the *minimax criterion* which is a special case of the α-value criterion where $\alpha = 0$. In the following section we define a measure of performance for policies with regard to the minimax criterion and give some results from theory of dynamic programming (DP).

4 DYNAMIC PROGRAMMING FOR THE MINIMAX CRITERION

4.1 THE VALUE FUNCTION

We call

$$\begin{aligned} V_\gamma^\pi(i) &:= \hat{m}_0\left(R_\gamma^\pi \mid I_0^\pi = i\right) \\ &= \sup\left\{r \in \mathrm{R} : P\left(R_\gamma^\pi > r \mid I_0^\pi = i\right) > 0\right\} \end{aligned}$$

the *max-value of the return of policy π with respect to state i*. The name max-value comes of the fact that $P\left(R_\gamma^\pi \leq V_\gamma^\pi(i) \mid I_0^\pi = i\right) = 1$ and $P\left(R_\gamma^\pi > r \mid I_0^\pi = i\right) > 0$ for every $r < V_\gamma^\pi(i)$. Hence, the agent which starts in state i and uses policy π can be sure that the total costs will not exceed $V_\gamma^\pi(i)$. But the agent has to be aware that the return will exceed any number less than $V_\gamma^\pi(i)$ with a positive probability. In other words, $V_\gamma^\pi(i)$ can be seen as the worst (i.e. maximum) value of the total costs that can occur if the agent uses π and starts in state i.

Imagine domains where it is to be guaranteed that the return will not exceed a given threshold r if the agent starts in state i. This constraint is satisfied if and only if a policy π is used that satisfies $V_\gamma^\pi(i) \leq r$. Examples for such domains are those goal tasks where the agent has to reach a goal *within a certain time* with probability one. The criterion of expectation is not adequate for these tasks in the framework of a MDP. This motivates the following alternative measure of performance for policies:

We measure the performance of a policy π by the *value function* V_γ^π which maps any $i \in S$ into $V_\gamma^\pi(i) \in \mathrm{R}$, and we define the *(minimax-)optimal value function* $V_\gamma^* \in \mathrm{R}^S$ by

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

We call a policy π *(minimax-)optimal* if its value function V_γ^π equals the minimax-optimal value function V_γ^*.

4.2 AN EXAMPLE

Before going on with the theory of the minimax criterion we give an example that reveals the utility of this criterion. In the MDP of figure 1, e.g.,

$$\begin{aligned} \mathrm{E}\left(R_\gamma^\pi \mid I_0^\pi = 0\right) &= 309 + 2\gamma; \qquad V_\gamma^\pi(0) = 400 + 2\gamma; \\ \mathrm{E}\left(R_\gamma^\mu \mid I_0^\mu = 0\right) &= 310 + 2\gamma; \qquad V_\gamma^\mu(0) = 310 + 2\gamma. \end{aligned}$$

Policy μ is minimax-optimal and ensures that the return will not exceed $310 + 2\gamma$ if the agent starts in state 0 whereas the worst-case return for π is $400 + 2\gamma$. Policy π is optimal with respect to the expected value criterion but there is only a difference of one cost unit for the expected return of μ. Therefore, policy π and μ have nearly the same performance with regard to the criterion of expectation. But with probability 0.75, the policy π will lead to a total cost of $400 + 2\gamma$ that is 90 units greater than the total cost of the minimax-optimal policy μ if the agent starts in state 0.

4.3 DYNAMIC PROGRAMMING ALGORITHMS

In (Heger, 1994), theory of DP for the minimax criterion is presented. Following is a short summary of the most essential results.

Let $N(i, a)$ be the set of states that immediately can be reached from state i by action a, i.e.,

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

for all $i \in S$ and $a \in A(i)$.

Furthermore, let $c(i, a, j)$ be the worst immediate cost that can occur in the transition from state i into state j under action a, i.e., to be precisely,

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

for all $i \in S$, $a \in A(i)$ and $j \in S$.

Theorem 1 *Let $\pi \in A^S$ be a stationary policy, $v_0 \in \mathrm{R}^S$ and*

$$v_{k+1}(i) := \max_{j \in N(i,\pi(i))} \left[c\left(i, \pi(i), j\right) + \gamma \cdot v_k(j)\right]$$

for all $i \in S$ and $k \in \mathrm{N}$. Then $v_k \in \mathrm{R}^S$ converges to V_γ^π as $k \to \infty$.

Theorem 2 *Let $v_0 \in \mathrm{R}^S$ and*

$$v_{k+1}(i) := \min_{a \in A(i)} \max_{j \in N(i,a)} \left[c\left(i, a, j\right) + \gamma \cdot v_k(j)\right]$$

for all $i \in S$ and $k \in \mathrm{N}$. Then $v_k \in \mathrm{R}^S$ converges to V_γ^* as $k \to \infty$.

Theorem 3 *A stationary policy $\pi \in A^S$ is optimal if and only if it satisfies*

$$\pi(i) = \arg\min_{a \in A(i)} \left(\max_{j \in N(i,a)} \left[c\left(i, a, j\right) + \gamma \cdot V_\gamma^*(j)\right]\right)$$

for all $i \in S$.

Theorem 1 yields a DP algorithm to compute the value function of a stationary policy π. By the algorithm of theorem 2, the optimal value function can be found. These algorithms use synchronous updating but we suspect that, by analogy with DP for the criterion of expectation (e.g. Barto, Bradtke & Singh, 1993; Williams & Baird, 1993), convergence holds for asynchronous kinds of DP as well.

Theorem 3 implies that there is at least one stationary optimal policy. From theorem 2 and 3 we see that all stationary optimal policies can be computed if the sets $N(i, a)$ and the worst immediate costs $c\left(i, a, j\right)$ are known for all $i, j \in S$ and $a \in A(i)$.

In contrast to DP for the criterion of expectation (Ross, 1970), it is not necessary to know complete probability distributions for state transitions. Therefore the model of the environment necessary in DP for the computation of minimax-optimal policies is much simpler than the corresponding one in DP for the criterion of expectation.

5 Q̂-LEARNING

We present the RL algorithm $\hat{Q}$-learning that finds worst-case optimal policies and can be regarded as a counterpart to Watkins' Q-learning (Watkins, 1989) that learns average-case optimal policies. The results presented here are proven in (Heger, 1994).

5.1 Q-NOTATION AND TASK SPECIFICATION

We define the *set of admissible state-action pairs* by

$$M := \{(i, a) : \ i \in S \text{ and } a \in A(i)\}.$$

Assume a function $Q \in \mathrm{R}^M$ and a state $i \in S$. We call an action $a \in A(i)$ *greedy with respect to i and Q* if and only if

$$a = \arg\min_{b \in A(i)} Q(i, b).$$

We call a stationary policy π *greedy with respect to Q* if and only if

$$\pi(i) = \arg\min_{a \in A(i)} Q(i, a)$$

for all $i \in S$.

Consider the *(minimax-)optimal Q-function* $\hat{Q}_\gamma^* \in \mathrm{R}^M$ that is defined by

$$\hat{Q}_\gamma^*(i, a) := \max_{j \in N(i,a)} \left[c(i, a, j) + \gamma \cdot V_\gamma^*(j)\right]$$

for all $i \in S$ and $a \in A(i)$. Theorem 3 implies that a stationary policy $\pi \in A^S$ is optimal if and only if it is greedy with respect to $\hat{Q}_\gamma^*$.

We call a function $Q \in \mathrm{R}^M$ *greedy-equivalent* to another function $Q' \in \mathrm{R}^M$ if and only if

$$\begin{aligned} &\{a \in A(i) : a \text{ is greedy with respect to } i \text{ and } Q\} \\ = \ &\{a \in A(i) : a \text{ is greedy with respect to } i \text{ and } Q'\} \end{aligned}$$

for all $i \in S$.

Assume that $Q \in \mathrm{R}^M$ is greedy-equivalent to $\hat{Q}_\gamma^*$. Again, theorem 3 implies that a stationary policy $\pi \in A^S$ is optimal if and only if it is greedy with respect to Q.

The task of $\hat{Q}$-learning (see figure 2) is to learn an arbitrary function $\hat{Q} \in \mathrm{R}^M$ that satisfies the condition that all stationary policies which are greedy with respect to $\hat{Q}$ are minimax-optimal. The reader should note that this task is not more difficult than the task to learn $\hat{Q}_\gamma^*$. It is even not more difficult than the task to find a function that is greedy-equivalent to $\hat{Q}_\gamma^*$.

5.2 CONDITIONS FOR CORRECTNESS

For the remainder of the text we assume that the Q-function in $\hat{Q}$-learning is represented by a table. We emphasize that all results in this section also hold if in the $\hat{Q}$-learning algorithm, the successor state of episode t differs from the starting state of episode $t + 1$.

The Q-function converges to a function $\hat{Q} \in \mathrm{R}^M$ in every case, no matter what actions will be selected by the agent. This can be shown easily because the Q-values are monotone increasing with respect to time and, at any time, they satisfy

$$\forall i \in S \ \forall a \in A(i): \ 0 \le q(i, a) \le Q(i, a) \le \frac{\sup C}{1 - \gamma}.$$

1. Initialize $Q(i,a) := q(i,a)$ for all $i \in S$ and $a \in A(i)$;
2. Let $i :=$ the starting state of the current episode;
3. Select an admissible action $a \in A(i)$ and execute it;
4. Let $j :=$ the successor state of the current episode;
5. Let $r :=$ the immediate cost of the current episode;
6. Let $Q(i,a) := \max\left[Q(i,a),\ r + \gamma \cdot \min_{b \in A(j)} Q(j,b)\right]$;
7. Go to Step 2.

Figure 2: The $\hat{Q}$-Learning Algorithm. The initial values $q(i,a)$ must satisfy $0 \leq q(i,a) \leq \hat{Q}^*_\gamma(i,a)$.

The Q-values even satisfy

$$\forall i \in S\ \forall a \in A(i):\ 0 \leq q(i,a) \leq Q(i,a) \leq \hat{Q}^*_\gamma(i,a)$$

at any time with probability one.

Therefore, the function $\hat{Q}$ satisfies

$$\forall i \in S\ \forall a \in A(i):\ 0 \leq q(i,a) \leq \hat{Q}(i,a) \leq \frac{\sup C}{1-\gamma};$$

and, with probability one,

$$\forall i \in S\ \forall a \in A(i):\ \hat{Q}(i,a) \leq \hat{Q}^*_\gamma(i,a).$$

The Q-function in $\hat{Q}$-learning converges to the optimal Q-function with probability one, i.e.,

$$\hat{Q} = \hat{Q}^*_\gamma$$

if the following condition holds: Every pair of starting state and action occurs infinitely often.

There is a definitely weaker condition that is sufficient to guarantee that $\hat{Q}$ is greedy-equivalent to $\hat{Q}^*_\gamma$. In order to explain this condition a further definition is helpful:

Let $\epsilon > 0$, $i \in S$, $a \in A(i)$ and $Q \in \mathrm{R}^M$. We say that a is *ϵ-greedy with respect to i and Q* if and only if

$$Q(i,a) - \min_{b \in A(i)} Q(i,b) \leq \epsilon.$$

Let $\epsilon > 0$ be a constant. Assume that every state becomes infinitely often a starting state. Further, assume that at any time the agent selects an action from the set of actions which are ϵ-greedy with respect to the current starting state and Q-function. Finally, assume that the agent always selects actions randomly and gives no ϵ-greedy action preference over another ϵ-greedy action, i.e. the probability that a is to be selected by the agent is the same for all ϵ-greedy actions a. This condition is still sufficient for $\hat{Q}$-learning to satisfy its task specification (section 5.1) because then, with probability one, $\hat{Q}$ is greedy-equivalent to $\hat{Q}^*_\gamma$ and, additionally,

$$\hat{Q}(i,a) = \hat{Q}^*_\gamma(i,a)$$

with probability one for every state $i \in S$ and action $a \in A(i)$ that is greedy with respect to i and $\hat{Q}^*_\gamma$.

Again, assume that every state becomes infinitely often a starting state. But now assume that the agent selects always a *greedy* action with respect to the current starting state and Q-function. It can be shown that under this condition the following equation holds with probability one for all $i \in S$:

$$\min_{a \in A(i)} \hat{Q}(i,a) = \min_{a \in A(i)} \hat{Q}^*_\gamma(i,a) = V^*_\gamma(i).$$

5.3 FIRST COMPARISONS WITH Q-LEARNING

In this section we suppose that the reader is familiar with Q-learning (Watkins, 1989; Watkins & Dayan 1992). Both $\hat{Q}$-learning and Q-learning are RL algorithms based on DP. But unlike the task of $\hat{Q}$-learning, Q-learning's task is to find a Q-function so that every stationary policy that is greedy with respect to this function is optimal with respect to the criterion of expectation. It is conspicuous that in $\hat{Q}$-learning there is no need for a learning rate and the Q-values are monotone increasing with respect to time. Both is generally not the case in Q-learning. Therefore, we suspect, that, generally, $\hat{Q}$-learning will converge faster than Q-learning.

The results of section 5.2 reveal that in in $\hat{Q}$-learning, the problem of the tradeoff between exploration and exploitation (e.g. Thrun, 1992) is smaller than in Q-learning. In both algorithms there is the problem that the agent has to visit all states infinitely often. If this condition can be secured the agent that uses $\hat{Q}$-learning may exploit always its current knowledge by choosing ϵ-greedy actions. On the other hand, in Q-learning the agent has to select *every* action infinitely often.

In deterministic domains $\hat{Q}$-learning identifies with the Q-learning algorithm if in the latter all Q-values are initialized as in Q-learning and if the learning rates in Q-learning are set to one. Hence, the complexity analysis of Q-Learning in determinstic domains (e.g. Koenig, 1992) holds for $\hat{Q}$-learning.

6 RELATED WORK

The modification of Q-learning presented in (Heger & Berns, 1992) is a preliminary algorithm of $\hat{Q}$-learning. It is also based on the minimax criterion but no convergence is ensured in probabilistic domains and it is not memoryless. It has been applied successfully in the coordination of the six legs of a simulated walking machine.

More mathematical details of the α-value criterion, the minimax criterion including DP and $\hat{Q}$-learning can be found in (Heger, 1994). DP algorithms with different notation and representation of uncertainty of the MDP for the minimax criterion also can be found in (Bertsekas & Rhodes, 1971) and (Witsenhausen, 1966).

7 CONCLUSIONS AND FUTURE WORK

This paper emphasizes the problems of the criterion of the expected return as a measure of the performance for policies. Despite the fact that a lot of work has been done in decision theory in order to obtain decision criteria that consider the phenomenon risk more sophisticatedly than the criterion of expectation, RL research is dominated by the expected value criterion until now. RL algorithms that are based on different decision criteria are still to be designed.

Especially in domains where it is to be ensured that the total costs will not exceed a threshold, the minimax criterion is to be preferred. The minimax criterion considers the worst-case of a decision's outcome and is to be chosen if security or the avoidance of risk becomes very important. $\hat{Q}$-learning is a RL algorithm with many pleasant features and is based on DP for the minimax criterion. Empirical tests of this algorithm are to be done in the future.

Acknowledgements

The author thanks Sven Koenig and Richard Yee for comments on this subject.

References

Barto, A. G., Bradtke, S. J. & Singh, S. P. (1993). Learning to Act using Real-Time Dynamic Programming. Department of Computer Science, Univ. of Massachusetts, Amherst, MA 01003. *Submitted to AI journal, special issue on Computational Theories of Interaction and Agency.*

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. & Rhodes, I. B. (1971), On the Minimax Feedback Control of Uncertain Systems. *Proceedings of the IEEE Decision and Control Conference, Miami, Dec, 1971*, pp. 451-455.

Bradley, J.V. (1976). *Probability; Decision; Statistics*. EnglewoodCliffs, New Jersey: Prentice Hall.

Heger, M. & Berns, K. (1992). Risikoloses Reinforcement-Lernen. *KI, Künstliche Intelligenz: Forschung, Entwicklung, Verfahren*, no 4, pp. 26-32, Organ des Fachbereichs 1 Künstliche Intelligenz der Gesellschaft für Informatik e.V. (GI).

Heger, M. (1994). Risk and Reinforcement Learning. *Technical report (forthcomming)*. Zentrum für Kognitionswissenschaften, Universität Bremen, Fachbereich 3 Informatik, Germany.

Koenig, S. (1992). The Complexity of Real-Time Reinforcement Learning Applied to Finding Shortest Paths in Deterministic Domains. *Technical Report CMU-CS-93-106*. Carnegie Mellon University, School of Computer Science. Pittsburgh, PA.

Koenig, S. & Simmons, R. G. (1993). Utility-Based Planning. *Technical report CMU-CS-93-222*. Carnegie Mellon University, School of Computer Science. Pittsburgh, PA.

von Neumann, J. & Morgenstern, O. (1947). *Theory of Games and Economic Behavior*. Princeton, New Jersey: Princeton Univ. Press.

Ross, S. M. (1970). *Applied Probability Models with Optimization Applications*. San Francisco, California: Holden Day.

Sutton, R. S. (1984). *Temporal Credit Assignment in Reinforcement Learning*. PhD thesis. Univ. of Massachusetts, Amherst, MA.

Sutton, R. S. (1991). Reinforcement Learning Architectures for Animats. In J. Meyers & S. Wilson (ed.), *Simulation of adaptive behavior: From animals to animats*, pp 288-296. Cambridge, MA: MIT Press.

Taha, H. A. (1987). *Operations Research: An Introduction, Fourth Edition*. New York: Macmillan Publishing Company.

Thrun, S. B. (1992). Efficient Exploration in Reinforcement Learning. *Technical report, CMU-CS-92-102*, School of Computer Science, Carnegie Mellon University. Pittsburgh, PA.

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). Analysis of Some Incremental Variants of Policy Iteration: First Steps Toward Understanding Actor-Critic Learning Systems. *Technical Report NU-CCS-93-11*. Northeastern University.

Witsenhausen, H.S. (1966). Minimax Controls of Uncertain Systems. *Technical Report ESL-R-269*. Department of Electrical Engineering, MIT. Cambridge, MA 02139.