The Effect of Representation and Knowledge on Goal-Directed Exploration with Reinforcement-Learning Algorithms: The Proofs

Sven Koenig and Reid G. Simmons

October 1995
CMU-CS-95-177

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

This research was supported in part by NASA under contract NAGW-1175 and by the "Studienstiftung des deutschen Volkes." The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the sponsoring organizations or the U.S. government.

Keywords: action models, admissible and consistent heuristics, action-penalty representation, complexity, goal-directed exploration, goal-reward representation, on-line reinforcement learning, prior knowledge, reward structure, Q-hat-learning, Q-learning

Abstract

This technical report contains the proofs of all theorems that are central to our Machine Learning Journal article "The Effect of Representation and Knowledge on Goal-Directed Exploration with Reinforcement-Learning Algorithms," in which we analyze the complexity of on-line reinforcement-learning algorithms that are applied to goal-directed exploration tasks: Previous work had concluded that, even in deterministic state spaces, initially uninformed reinforcement learning was at least exponential for such problems, or that it was of polynomial worst-case time-complexity only if the learning methods were augmented. In the article we prove that, to the contrary, the algorithms are tractable with only a simple change in the reward structure ("penalizing the agent for action executions") or in the initialization of the values that they maintain. In particular, we provide tight complexity bounds for both Watkins' Q-learning and Heger's Q-hat-learning and show how their complexity depends on properties of the state spaces. We also demonstrate how one can decrease the complexity even further by either learning action models or utilizing prior knowledge of the topology of the state spaces. Our results provide guidance for empirical reinforcement-learning researchers on how to distinguish hard reinforcement-learning problems from easy ones and how to represent them in a way that allows them to be solved efficiently.

This technical report contains the proofs of the theorems used in our Machine Learning Journal article "The Effect of Representation and Knowledge on Goal-Directed Exploration with Reinforcement-Learning Algorithms" that we have not already proved there. The report uses the terminology and definitions described in that article.

We assume that all state spaces are safely explorable and (without loss of generality) that all states that cannot be reached from the start state or that can only be reached by passing through a goal state have been deleted.

First, we prove several properties of consistent Q-values, for example how they relate to admissible Q-values. We prove properties of admissible Q- or $\hat{\text{Q}}$-learning in deterministic state spaces if the initial Q-values are consistent. Then, we show that the same properties hold for admissible $\hat{\text{Q}}$-learning in non-deterministic state spaces (which include deterministic state spaces as a special case) if the initial Q-values are admissible. Finally, we prove our complexity results.

All proofs are for the undiscounted case with action-penalty representation. The transformations of the Q-values that we have described in Sections 6.1.1.2. or 6.1.2. of the article can then be used to transfer the results to discounted Q- or $\hat{\text{Q}}$-learning with action-penalty representation or goal-reward representation, respectively.

The time superscripts t used in the following refer to the values of the variables immediately before the agent executes the $(t+1)$st action.

Theorem 1 *If Q-values are consistent for undiscounted Q- or $\hat{Q}$-learning with action-penalty representation, then* $-\min_{s' \in X} d(s,s') \leq U(s) \leq 0$ *for all* $s \in S$*, where* $X := \{s \in S : U(s) = 0\}$.

Proof by induction on $d'(s) := \min_{s' \in X} d(s,s')$. Note that $d'(s)$ is finite for all $s \in S$, since the state space is safely explorable and $G \subseteq X$.

- If $d'(s) = 0$, then $s \in X$ and therefore $U(s) = 0$. Thus, $-d'(s) = U(s) = 0$.
- If $d'(s) \neq 0$, then $s \notin X \supseteq G$ and therefore $s \in S \setminus G$. Let $a := \text{argmin}_{a' \in A(s)} \max_{s' \in succ(s,a')} d'(s')$. Assume that the theorem holds for all $s'' \in S$ with $d'(s'') < d'(s)$. Then, it holds for all $s''' \in succ(s,a)$, since $\max_{s' \in succ(s,a)} d'(s') = d'(s) - 1 < d'(s)$ and therefore $d'(s''') < d'(s)$. Thus, $-d'(s) = -1 - \max_{s' \in succ(s,a)} d'(s') \overset{\text{Assumption}}{\leq} -1 + \min_{s' \in succ(s,a)} U(s') \overset{\text{Consistency}}{\leq} Q(s,a) \leq \max_{a' \in A(s)} Q(s,a') = U(s) \leq \max_{s' \in S, a' \in A(s)} Q(s',a') \overset{\text{Consistency}}{\leq} 0$.

Theorem 2 *If Q-values are consistent for undiscounted Q- or $\hat{Q}$-learning with action-penalty representation, then* $-1 - \max_{s' \in succ(s,a)} \min_{s'' \in X} d(s',s'') \leq Q(s,a) \leq 0$ *for all* $s \in S \setminus G$ *and* $a \in A(s)$*, where* $X := \{s \in S : U(s) = 0\}$.

Proof: $-1 - \max_{s' \in succ(s,a)} \min_{s'' \in X} d(s', s'') \overset{\text{Theorem 1}}{\leq} -1 + \min_{s' \in succ(s,a)} U(s') \overset{\text{Consistency}}{\leq} Q(s,a) \overset{\text{Consistency}}{\leq} 0$ for all $s \in S \setminus G$ and $a \in A(s)$.

Theorem 3 *Consistent Q-values for undiscounted Q- or $\hat{Q}$-learning with action-penalty representation are admissible for undiscounted Q- or $\hat{Q}$-learning with action-penalty representation.*

Proof: Assume consistent Q-values. $Q(s,a) = 0$ for all $s \in G$ and $a \in A(s)$. It holds for all $s \in S \setminus G$ and $a \in A(s)$ that $-1 - \max_{s' \in succ(s,a)} gd(s') = -1 - \max_{s' \in succ(s,a)} \min_{s'' \in G} d(s', s'') \leq -1 - \max_{s' \in succ(s,a)} \min_{s'' \in X} d(s', s'') \overset{\text{Theorem 1}}{\leq} -1 + \min_{s' \in succ(s,a)} U(s') \overset{\text{Consistency}}{\leq} Q(s,a) \overset{\text{Consistency}}{\leq} 0$, where $G \subseteq X := \{s \in S : U(s) = 0\}$.

Theorem 4 *If Q-values are admissible for undiscounted Q- or $\hat{Q}$-learning with action-penalty representation, then $-gd(s) \leq U(s) \leq 0$ for all $s \in S$.*

Proof: If $s \in G$, then $-gd(s) = 0 = \max_{a \in A(s)} 0 \overset{\text{Admissibility}}{=} \max_{a \in A(s)} Q(s,a) = U(s)$. It holds for all $s \in S \setminus G$ that $-gd(s) = -(1 + \min_{a \in A(s)} \max_{s' \in succ(s,a)} gd(s')) = \max_{a \in A(s)}(-1 - \max_{s' \in succ(s,a)} gd(s')) \overset{\text{Admissibility}}{\leq} \max_{a \in A(s)} Q(s,a) = U(s) \leq \max_{s' \in S, a' \in A(s)} Q(s', a') \overset{\text{Admissibility}}{\leq} 0$.

Consider the following algorithm **Alg1** in a deterministic state space: Given arbitrary Q-values, pick an arbitrary state $s \in S \setminus G$ and determine $a := \text{argmax}_{a' \in A(s)} Q(s, a')$. (Ties can be broken arbitrarily.) Let s' be the uniquely determined successor state if a is executed in s, i.e. $s' = succ(s,a)$. Set $Q(s,a) := -1 + U(s')$ and leave the other Q-values unchanged. Refer to the old Q-values as $Q^0(s,a)$ and to the new ones as $Q^1(s,a)$, i.e.

$$Q^1(s'', a') = \begin{cases} -1 + U^0(s') & \text{if } s'' = s \text{ and } a' = a \\ Q^0(s'', a') & \text{otherwise} \end{cases} \qquad \text{for all } s'' \in S \text{ and } a' \in A(s'')$$

(Note that the value-update step of algorithm Alg1 is the value-update step used in the first part of the definition of undiscounted admissible Q- or $\hat{\text{Q}}$-learning with action-penalty representation in deterministic state spaces.)

Theorem 5 *If the Q^0-values are consistent for undiscounted Q- or $\hat{Q}$-learning with action-penalty representation, then*

1. $Q^1(s'', a') \leq Q^0(s'', a')$ *for all* $s'' \in S$ *and* $a' \in A(s'')$,

2. *$U^1(s'') \leq U^0(s'')$ for all $s'' \in S$, and*
3. *the Q^1-values are consistent for undiscounted Q- or $\hat{\text{Q}}$-learning with action-penalty representation.*

Proof:

1. $Q^1(s'',a') = -1 + U^0(s') = -1 + \min_{s''' \in succ(s'',a')} U^0(s''') \overset{\text{Consistency}}{\leq} Q^0(s'',a')$ for $s'' = s$ and $a' = a$, and $Q^1(s'',a') = Q^0(s'',a')$ otherwise. Thus, $Q^1(s'',a') \leq Q^0(s'',a')$ for all $s'' \in S$ and $a' \in A(s'')$.
2. According to the first part of this theorem, it holds that $Q^1(s'',a') \leq Q^0(s'',a') \overset{\text{Consistency}}{\leq} 0$ for all $s'' \in S$ and $a' \in A(s'')$. Then, $U^1(s'') = \max_{a' \in A(s'')} Q^1(s'',a') \leq \max_{a' \in A(s'')} Q^0(s'',a') = U^0(s'')$ for all $s'' \in S$.
3. According to the second part of this theorem, it holds that $U^1(s'') \leq U^0(s'')$ for all $s'' \in S$. Then, $-1 + \min_{s''' \in succ(s'',a')} U^1(s''') = -1 + U^1(s') \leq -1 + U^0(s') = Q^1(s'',a') \leq 0$ for $s'' = s$ and $a' = a$, $Q^1(s'',a') = Q^0(s'',a') = 0$ for all $s'' \in G$ and $a' \in A(s'')$, and $-1 + U^1(succ(s'',a')) \leq -1 + U^0(succ(s'',a')) \overset{\text{Consistency}}{\leq} Q^0(s'',a') = Q^1(s'',a') = Q^0(s'',a') \overset{\text{Consistency}}{\leq} 0$ otherwise. Thus, the Q^1-values are consistent for undiscounted Q- or $\hat{\text{Q}}$-learning with action-penalty representation.

Theorem 6 *If the initial Q-values of algorithm Alg1 (see Theorem 5) in a deterministic state space are consistent for undiscounted Q- or $\hat{\text{Q}}$-learning with action-penalty representation, then they remain consistent after every action execution, and the Q-values and U-values are monotonically decreasing.*

Proof by induction on the number of action executions: The Q-values are consistent before the first action execution. Assume that they are consistent before an arbitrary action execution. According to Theorem 5, they are consistent after the action execution, and the Q-values and U-values are monotonically decreasing.

Consider the following algorithm **Alg2** in a non-deterministic state space: Given arbitrary Q-values, pick an arbitrary state $s \in S \setminus G$ and determine $a := \text{argmax}_{a' \in A(s)} Q(s,a')$. (Ties can be broken arbitrarily.) Let s' be an arbitrary successor state if a is executed in s, i.e. $s' \in succ(s,a)$. Set $Q(s,a) := \min(Q(s,a), -1 + U(s'))$ and leave the other Q-values unchanged. Refer to the old Q-values as $Q^0(s,a)$ and to the new ones as $Q^1(s,a)$, i.e.

$$Q^1(s'',a') = \begin{cases} \min(Q^0(s'',a'), -1 + U^0(s')) & \text{if } s'' = s \text{ and } a' = a \\ Q^0(s'',a') & \text{otherwise} \end{cases} \quad \text{for all } s'' \in S \text{ and } a' \in A(s'')$$

(Note that the value-update step of algorithm Alg2 is the value-update step used in the definition of undiscounted admissible $\hat{\text{Q}}$-learning with action-penalty representation in non-deterministic state spaces, of which the value-update step of the second part of the definition of undiscounted admissible Q- or $\hat{\text{Q}}$-learning with action-penalty representation in deterministic state spaces is a special case.)

Theorem 7 *If the Q^0-values are admissible, then*

1. *$Q^1(s'', a') \le Q^0(s'', a')$ for all $s'' \in S$ and $a' \in A(s'')$,*
2. *$U^1(s'') \le U^0(s'')$ for all $s'' \in S$, and*
3. *the Q^1-values are admissible for undiscounted Q- or $\hat{\text{Q}}$-learning with action-penalty representation.*

Proof:

1. $Q^1(s'', a') = \min(Q^0(s'', a'), -1 + U^0(s')) \le Q^0(s'', a')$ for $s'' = s$ and $a' = a$, and $Q^1(s'', a') = Q^0(s'', a')$ otherwise. Thus, $Q^1(s'', a') \le Q^0(s'', a')$ for all $s'' \in S$ and $a' \in A(s'')$.
2. According to the first part of this theorem, it holds that $Q^1(s'', a') \le Q^0(s'', a') \overset{\text{Admissibility}}{\le} 0$ for all $s'' \in S$ and $a' \in A(s'')$. Then, $U^1(s'') = \max_{a' \in A(s'')} Q^1(s'', a') \le \max_{a' \in A(s'')} Q^0(s'', a') = U^0(s'')$ for all $s'' \in S$.
3. $-1 - \max_{s''' \in succ(s'', a')} gd(s''') \le Q^0(s'', a')$ and $-1 - \max_{s''' \in succ(s'', a')} gd(s''') \le -1 - gd(s') \overset{\text{Theorem 4}}{\le} -1 + U^0(s')$ for $s'' = s$ and $a' = a$, therefore $-1 - \max_{s''' \in succ(s'', a')} gd(s''') \le \min(Q^0(s'', a'), -1 + U^0(s')) = Q^1(s'', a') \le 0$ for $s'' = s$ and $a' = a$. $Q^1(s'', a') = Q^0(s'', a') = 0$ for all $s'' \in G$ and $a' \in A(s'')$, and $-1 - \max_{s''' \in succ(s'', a')} gd(s''') \overset{\text{Admissibility}}{\le} Q^0(s'', a') = Q^1(s'', a') = Q^0(s'', a') \overset{\text{Admissibility}}{\le} 0$ otherwise. Thus, the Q^1-values are admissible for undiscounted Q- or $\hat{\text{Q}}$-learning with action-penalty representation.

Theorem 8 *If the initial Q-values of algorithm Alg2 (see Theorem 7) in a non-deterministic state space are admissible for undiscounted Q- or $\hat{\text{Q}}$-learning with action-penalty representation, then they remain admissible after every action execution, and the Q-values and U-values are monotonically decreasing.*

Proof by induction on the number of action executions: The Q-values are admissible before the first action execution. Assume that they are admissible before an arbitrary action execution. According to Theorem 7, they are admissible after the action execution, and the Q-values and U-values are monotonically decreasing.

Theorem 9 *For all $t \in \mathcal{N}_0$ (until termination) of (a) undiscounted admissible Q- or $\hat{Q}$-learning with action-penalty representation in deterministic state spaces and (b) undiscounted admissible $\hat{Q}$-learning with action-penalty representation in non-deterministic state spaces, it holds that $U^t(s^t) + \sum_{s \in S} \sum_{a \in A(s)} Q^0(s,a) - t \geq \sum_{s \in S} \sum_{a \in A(s)} Q^t(s,a) + U^0(s^0) - loop^t$ and $loop^t \leq \sum_{s \in S} \sum_{a \in A(s)} Q^0(s,a) - \sum_{s \in S} \sum_{a \in A(s)} Q^t(s,a)$, where $loop^t := |\{t' \in \{0, \ldots, t-1\} : s^{t'} = s^{t'+1}\}|$ (the number of actions executed before t that did not change the state).*

Proof by induction on t: The theorem trivially holds for $t = 0$. Assume that it holds for an arbitrary t. Note that $Q^t(s^t, a^t) = U^t(s^t)$, due to the specific action-selection step used. We distinguish two cases:

- The action executed at t did not change the state:

  Then, $s^{t+1} = s^t$ and $loop^{t+1} = 1 + loop^t$. Depending on the value-update step used, it holds that either $Q^{t+1}(s^t, a^t) = -1 + U^t(s^{t+1}) = -1 + U^t(s^t) = -1 + Q^t(s^t, a^t)$ (for the first part of the definition of admissible Q- or $\hat{\text{Q}}$-learning in deterministic state spaces) or $Q^{t+1}(s^t, a^t) = \min(Q^t(s^t, a^t), -1 + U^t(s^t)) = \min(Q^t(s^t, a^t), -1 + Q^t(s^t, a^t)) = -1 + Q^t(s^t, a^t)$ (otherwise). Thus, in both cases $Q^{t+1}(s^t, a^t) = -1 + Q^t(s^t, a^t)$. $U^{t+1}(s^{t+1}) = U^{t+1}(s^t) = \max_{a \in A(s^t)} Q^{t+1}(s^t, a) \geq Q^{t+1}(s^t, a^t) = -1 + Q^t(s^t, a^t) = -1 + U^t(s^t)$. All other values do not change from t to $t+1$.

  $U^{t+1}(s^{t+1}) + \sum_{s \in S} \sum_{a \in A(s)} Q^0(s,a) - (t+1) \geq (-1 + U^t(s^t)) + \sum_{s \in S} \sum_{a \in A(s)} Q^0(s,a) - (t+1) = (U^t(s^t) + \sum_{s \in S} \sum_{a \in A(s)} Q^0(s,a) - t) - 2 \overset{\text{Assumption}}{\geq} (\sum_{s \in S} \sum_{a \in A(s)} Q^t(s,a) + U^0(s^0) - loop^t) - 2 = (-1 + \sum_{s \in S} \sum_{a \in A(s)} Q^t(s,a)) + U^0(s^0) - (1 + loop^t) = \sum_{s \in S} \sum_{a \in A(s)} Q^{t+1}(s,a) + U^0(s^0) - loop^{t+1}$.

  $loop^{t+1} = 1 + loop^t \overset{\text{Assumption}}{\leq} 1 + (\sum_{s \in S} \sum_{a \in A(s)} Q^0(s,a) - \sum_{s \in S} \sum_{a \in A(s)} Q^t(s,a)) = \sum_{s \in S} \sum_{a \in A(s)} Q^0(s,a) - (-1 + \sum_{s \in S} \sum_{a \in A(s)} Q^t(s,a)) = \sum_{s \in S} \sum_{a \in A(s)} Q^0(s,a) - \sum_{s \in S} \sum_{a \in A(s)} Q^{t+1}(s,a)$.

  In other words, the theorem also holds for $t+1$.

- The action executed at t changed the state:

  Then, $s^{t+1} \neq s^t$, $loop^{t+1} = loop^t$, and (for both possible value-update steps) $Q^{t+1}(s^t, a^t) \leq -1 + U^t(s^{t+1}) = -1 + U^{t+1}(s^{t+1})$. All other values, except for $U(s^t)$, do not change from t to $t+1$.

  $U^{t+1}(s^{t+1}) + \sum_{s \in S} \sum_{a \in A(s)} Q^0(s,a) - (t+1) = (U^t(s^t) + \sum_{s \in S} \sum_{a \in A(s)} Q^0(s,a) - t) + U^{t+1}(s^{t+1}) - U^t(s^t) - 1 \overset{\text{Assumption}}{\geq} (\sum_{s \in S} \sum_{a \in A(s)} Q^t(s,a) + U^0(s^0) - loop^t) + U^{t+1}(s^{t+1}) - U^t(s^t) - 1 = (-1 + U^{t+1}(s^{t+1})) - U^t(s^t) + \sum_{s \in S} \sum_{a \in A(s)} Q^t(s,a) + U^0(s^0) - loop^t \geq (Q^{t+1}(s^t, a^t) - Q^t(s^t, a^t) + \sum_{s \in S} \sum_{a \in A(s)} Q^t(s,a)) + U^0(s^0) - loop^t = (\sum_{s \in S} \sum_{a \in A(s)} Q^{t+1}(s,a)) + U^0(s^0) - loop^{t+1}$.

  $loop^{t+1} = loop^t \overset{\text{Assumption}}{\leq} \sum_{s \in S} \sum_{a \in A(s)} Q^0(s,a) - \sum_{s \in S} \sum_{a \in A(s)} Q^t(s,a) \leq \sum_{s \in S} \sum_{a \in A(s)} Q^0(s,a) - \sum_{s \in S} \sum_{a \in A(s)} Q^{t+1}(s,a)$, since $Q^{t+1}(s,a) \leq Q^t(s,a)$ for all $s \in S$ and $a \in A(s)$ according to Theorems 6 and 8.

In other words, the theorem also holds for $t+1$.

Theorem 10 *(A) Undiscounted admissible Q- or $\hat{Q}$-learning with action-penalty representation in deterministic state spaces and (b) undiscounted admissible $\hat{Q}$-learning with action-penalty representation in non-deterministic state spaces reach a goal state after at most $2\sum_{s\in S\setminus G}\sum_{a\in A(s)}(Q^0(s,a)+\max_{s'\in succ(s,a)} gd(s')+1) - U^0(s^0)$ action executions in non-deterministic state spaces.*

Proof: $t \overset{\text{Theorem 9}}{\leq} U^t(s^t) + \sum_{s\in S}\sum_{a\in A(s)} Q^0(s,a) - \sum_{s\in S}\sum_{a\in A(s)} Q^t(s,a) - U^0(s^0) + loop^t \overset{\text{Theorem 9}}{\leq} U^t(s^t) + \sum_{s\in S}\sum_{a\in A(s)} Q^0(s,a) - \sum_{s\in S}\sum_{a\in A(s)} Q^t(s,a) - U^0(s^0) + (\sum_{s\in S}\sum_{a\in A(s)} Q^0(s,a) - \sum_{s\in S}\sum_{a\in A(s)} Q^t(s,a)) = U^t(s^t) + 2\sum_{s\in S}\sum_{a\in A(s)} Q^0(s,a) + 2\sum_{s\in S}\sum_{a\in A(s)} -Q^t(s,a) - U^0(s^0) = U^t(s^t) - U^0(s^0) + 2\sum_{s\in S}\sum_{a\in A(s)}(Q^0(s,a) - Q^t(s,a)) \leq -U^0(s^0) + 2\sum_{s\in S\setminus G}\sum_{a\in A(s)}(Q^0(s,a)+\max_{s'\in succ(s,a)} gd(s')+1)$, since the Q-values are admissible according to Theorems 6 and 3 or Theorem 8, and therefore $Q^0(s,a) = Q^t(s,a) = 0$ for all $s\in G$ and $a\in A(s)$, $-1-\max_{s'\in succ(s,a)} gd(s') \overset{\text{Admissibility}}{\leq} Q^t(s,a)$ for all $s\in S\setminus G$ and $a\in A(s)$, and $U^t(s) \overset{\text{Theorem 4}}{\leq} 0$ for all $s\in S$.

Theorem 11 *(A) Undiscounted admissible Q- or $\hat{Q}$-learning with action-penalty representation in deterministic state spaces and (b) undiscounted admissible $\hat{Q}$-learning with action-penalty representation in non-deterministic state spaces reach a goal state after at most $O(ed)$ action executions in non-deterministic state spaces.*

Proof: The algorithm reaches a goal state after at most $O(2\sum_{s\in S\setminus G}\sum_{a\in A(s)}(Q^0(s,a)+\max_{s'\in succ(s,a)} gd(s')+1) - U^0(s^0)) \leq O(2\sum_{s\in S\setminus G}\sum_{a\in A(s)}(d+1)+d) \leq O(2e(d+1)+d) = O(ed)$ action executions according to Theorem 10, since the Q-values are admissible according to Theorems 6 and 3 or Theorem 8, and therefore $Q^0(s,a) \overset{\text{Admissibility}}{\leq} 0$ for all $s\in S$ and $a\in A(s)$, and $-d \leq -gd(s) \overset{\text{Theorem 4}}{\leq} U^0(s)$ for all $s\in S$.

Thanks to Diana Gordon for pointing out a typo in the original manuscript.