UNIVERSITEIT VAN AMSTERDAM
Faculteit Natuurwetenschappen, Wiskunde en Informatica

Reinforcement Learning Week 1

By

Sergio Alejandro Gutierrez Maury, 11821353
Max Emanuel Feucht, 13620568

September 13, 2023

1 Question 2.4: Coding Assignment - Dynamic Programming

In the lab you implemented the value iteration and policy iteration algorithm.

- For which of these algorithms do you expect a single iteration to run faster? Briefly explain why. (For policy iteration, one iteration includes the policy evaluation and policy improvement steps).

  Answer: A single iteration of the Value Iteration algorithm is expected to run faster because it combines both policy evaluation and policy improvement into a single update operation, whereas in Policy Iteration, each iteration involves multiple sweeps (policy evaluation and improvement steps).

- Which of the algorithms do you expect to take fewer total iterations? Briefly explain why.

  Answer: Policy Iteration is expected to take fewer total iterations to converge to an optimal policy. This is because Policy Iteration explicitly separates the policy evaluation and policy improvement steps, allowing for quicker convergence within each iteration. In contrast, Value Iteration requires more iterations as it performs both policy evaluation and improvement simultaneously in each iteration, making it slower to converge.

2 Question 2.5: Dynamic Programming

1 Write the value, $v_\pi(s)$, of a state s under policy π, in terms of π and $q_\pi(s,a)$. Write down both the stochastic and the deterministic policy case.

  Answer:

  For the deterministic case, $v_\pi(s)$ can be written as $\sum_a q_\pi(s,a)\pi(s)$, while for the stochastic case $v_\pi(s) = \sum_a q_\pi(s,a)\pi(a|s)$. Note the difference between $\pi(s)$ and $\pi(a|s)$.

2 The Value Iteration update, given in the book in Eq. 4.10 can also be rewritten in terms of Q-values. Give the Q-value Iteration update. Eq.10 states that $v_{k+1}(s) = \max_a \sum_{s',r} p(s',r|s,a)(r+\gamma v_k(s))$, while we can express $q_k(s,a)$ in nearly the exact same form, i.e., $q_k(s,a) = \sum_{s',r} p(s',r|s,a)(r+\gamma v_k(s))$, only without the *max* argument.

From this follows that:

$$v_{k+1}(s) = \max_a q_k(s,a) \tag{1}$$

$$q_{k+1}(s) = \sum_{s',r} p(s',r|s,a)(r+\gamma v_{k+1}(s')) \tag{2}$$

$$q_{k+1}(s) = \sum_{s',r} p(s',r|s,a)(r+\gamma \max_{a'} q_k(s',a')) \tag{3}$$

3 In Policy Iteration, we first evaluate the policy by computing its value function, and then update it using a Policy Improvement step. You will now change the Policy Evaluation algorithm on page 74 of RL:AI to compute action values. Give the new policy evaluation update in terms of $Q_\pi(s,a)$ instead of $V_\pi(s)$. That is, write an update for $Q_\pi(s,a)$ in terms of the current action-value function. Your answer should not contain a state-value function.

Answer: Based on the formulas provided in 3, the update for $Q(s,a)$ looks as follows:

$$Q(s,a) = \sum_{s',r} p(s',r|s,a)(r+\gamma \max_{a'} Q(s',a')) \tag{4}$$

4 Now change the Policy Improvement step of the algorithm on page 80 of RL:AI in terms of $Q^\pi(s,a)$ instead of $V^\pi(s)$.

Answer: The Policy Improvement algorithm becomes simpler when formulating it in terms of $Q(s,a)$:

Algorithm 1 Policy Improvement

```
$policy\_stable \leftarrow true$
for $s \in S$ do
    $old\_action \leftarrow \pi(s)$
    $\pi(s) \leftarrow \arg\max_a Q(s,a))$
    if $old\_action \neq \pi(s)$ then $policy\_stable \leftarrow false$
    end if
end for
if $policy\_stable$ then stop and return $V \approx v_*$ and $\pi \approx \pi_*$, else Policy Evaluation
end if
```

This formulation makes intuitive sense, as 1) the formula for updating $\pi(s)$ on page 80 after the arg max equals exactly the formula for $Q(s,a)$ when formulated in terms of $V(s')$, and as 2) $(Q(s,a)$ evaluates state-action pairs in combination automatically.

5 You might have noticed the policy evaluation step on page 80 is different from the separate algorithm on page 75. What is the difference, and why do you think this difference exists?

Answer:

The update value function on page 75 is:

$$V(s) = \underbrace{\sum_{a} \pi(a|s) \cdot}_{} \sum_{s',r} p(s',r|s,\pi(s))[r+\gamma \cdot V(s')]$$

And the update value function from page 80 is:

$$V(s) = \sum_{s',r} p(s',r|s,\pi(s))[r+\gamma \cdot V(s')]$$

On page 75, the algorithm uses the stochastic version of the Bellman equation, because it explicitly considers the action probabilities $\pi(a|s)$ when updating the value function V(s). It accounts for the uncertainty associated with the policy's choice of actions.

On page 80, the algorithm aims to estimate V(s) for a specific deterministic policy, which means that each state for each s, there is only one action chosen with certainty, this is done in the policy improvement step, with:

$$\pi(s) = argmax_a \sum_{s',r} p(s',r|s,\pi(s))[r+\gamma \cdot V(s')]$$