Homework 1

Elias Dubbeldam, Ankur Satya

September 16, 2022

2.4 Coding Assignment - Dynamic Programming

1. See `codegra.de` for the notebook.

2. We have implemented the value iteration and policy iteration algorithms in the lab.

   Policy iteration cycles between the two steps of policy evaluation and policy improvement. In both steps, once has to loop over all states s. Each policy evaluation step stops once there is convergence.

   Value iteration cycles implicitly between two steps. The policy is evaluated only once for each s, where after the policy is updated directly. Hence, the policy evaluation step is similar as in policy iteration. However, it does not wait until the convergence before it updates (improves) the policy.

   - For a single iteration, the policy evaluation step has to converge for policy iteration before a new iteration can be started. This is not the case for value iteration, it directly continues after one update for each s. Therefore, we expect that a single value iteration is faster than a single policy iteration
   - Policy iteration takes fewer iterations in total. The policy is updated once the policy evaluation step has converged, which makes the policy improvement 'more effective' for an individual iteration. The policy update in value iteration happens 'less informed', because it happens directly after one update during the iteration, and does not wait until convergence.

2.5 Dynamic Programming

See the written notes below.

2.5. Dynamic Programming

(1) Stochastic:

using eq^n 4.4, we can say:

$$v_\pi(s) = E_\pi[G_t \mid S_t = s]$$
$$= \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)[r + \gamma v_\pi(s')]$$

Using eq^n 4.6, we can say

$$q_\pi(s,a) = E[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s, A_t = a]$$
$$= \sum_{s',r} p(s',r|s,a)[r + \gamma v_\pi(s')]$$

Using the above two eqn^s, we get:-

$$v_\pi(s) = \sum_a \pi(a|s)\, q_\pi(s,a)$$

~~(2)~~ Deterministic :-

$$v^\pi(s) = q_\pi(s,a) \ \forall\, a \in A: \pi(a|s) = 1$$

(2) The Bellman optimality eq^n for the action-value function:

$$q_*(s,a) = \sum_{s',r} p(s',r|s,a)\left[r + \gamma \max_{a'} q_*(s',a')\right].$$

The value-iteration for the state-value function was obtained by turning its Bellman eq^n to an update rule. [BartoSutton] using the same for the action-value ~~function~~ iteration:-

$$q_{k+1}(s) = \sum_{s'} p(s', r \mid s, a)\left[r + \gamma \max_{a'} q_k(s', a')\right].$$

(3) We know:

$$v_\pi(s) = \sum_{a \in A} \pi(a \mid s) \cdot q_\pi(s, a) \quad -① $$

$$q_\pi(s) = \sum_{s'} p(s', r \mid s, a)\left[r + \gamma v_\pi(s')\right]. \quad -②$$

Replacing $v_\pi(s)$ from ① to ②

$$q_\pi(s) = \sum_{s'} p(s', r \mid s, a)[r + \gamma \sum_{a' \in A} \pi(a' \mid s) \cdot q_\pi(s', a')$$

~~(4) Using eqn 4.9, we can say~~

$\cancel{\pi}$

(4) Given:

$$\pi(s) \leftarrow \arg\max_{a} \sum_{s',r} p(s',r|s,a)[r + \gamma V(s')]$$

We also know that:-

$$v_\pi(s) = \sum_{a \in A} \pi(a|s) \cdot q_\pi(s,a) \quad - (1)$$

Using (1) we can rewrite the policy improvement step:-

$$\pi(s) \leftarrow \arg\max_{a} \sum_{s',r} p(s',r|s,a)\left[r + \gamma \sum_{a' \in A} \pi(a'|s) \cdot q_\pi(s,a')\right]$$

(5) The policy evaluation step on page 75 assumes that $\pi(s)$ has a probability distribution over 'a' $\in A$. That's why a summation over $\forall\, a \in A$ is taken on page 75.

The evaluation step on page 80 assumes a deterministic or a pure greedy policy with respect to the actions. Hence, there is only one action that should be chosen for a state.