Reinforcement Learning

Homework Chapters 1, 2

Tom Lieberum - ID 13253042

2.1: Dynamic Programming

1. Stochastic:

$$v^\pi(s) = \mathop{\mathbb{E}}_{a\sim\pi(\cdot|s)}[q^\pi(s,a)]$$

$$= \begin{cases} \sum_{a\in\mathcal{A}} \pi(a|s)\cdot q^\pi(s,a), \text{ discrete case} \\ \int_{\mathcal{A}} \pi(a|s)\cdot q^\pi(s,a)\, da, \text{ continuous case} \end{cases}$$

   Deterministic:

$$v^\pi(s) = q^\pi(s,\pi(s))$$

2.

$$q_{k+1}(s,a) = \mathbb{E}\left[R_{t+1} + \gamma \max_{a'\in\mathcal{A}} q_k(S_{t+1},a') | S_t = s, A_t = a\right]$$

$$= \sum_{s',r} p(s',r|s,a)\left[r + \gamma \max_{a'\in\mathcal{A}} q_k(s',a')\right]$$

3.

$$Q^\pi_{k+1} = \sum_{s',r} p(s',r|s,a)\left[r + \gamma \sum_{a'\in\mathcal{A}} \pi(a'|s')\cdot Q^\pi_k(s',a')\right]$$

4.

$$\pi_{new}(s) = \operatorname*{argmax}_{a\in\mathcal{A}} Q^{\pi_{old}}(s,a)$$

2.2: Coding Assignment - Dynamic Programming

1. Cf. code.
2. During each iteration of value iteration we compute the maximum over $|\mathcal{A}|$ elements which take themselves $O(|\mathcal{S}||\mathcal{R}|)$ operations to compute. So each iteration in VI has time complexity $O(|\mathcal{S}||\mathcal{R}||\mathcal{A}|)$, if we assume that evaluating v has constant time complexity (e.g. when we are in a tabular case).

In policy iteration, we have two steps per iteration. In the evaluation step, each improvement in the evaluation costs $O(|\mathcal{S}||\mathcal{R}||\mathcal{A}|)$ time, which we need to multiply by the number of iterations it takes for the evaluation to converge to v^π.

In the improvement step, we essentially perform the same computation as in the value iteration update, just that we are concerned with the argmax, rather than the max. So this step costs $O(|\mathcal{S}||\mathcal{R}||\mathcal{A}|)$ as well.

So, if we assume that it takes value iteration a factor of α more global iterations to converge than policy iteration (e.g. $\alpha = 1$), then value iteration is faster by a factor of N/α where N is the expected number of iterations we need for the policy evaluation step to converge to v^π.