1.6 Summary

In chapter 1. we proposed following concepts:

① Discounted MDPs

② Finite-Horizon MDPs

③ Policy and iteration optimization.

we proved following significant theorems:

① [Th 1.7] Bellman Optimality Equations:

$$V^* = V^{\pi^*}, \quad Q^* = Q^{\pi^*}$$

② [Th 1.8] Bellman Optimality Equations:

$$Q(s,a) = r(s,a) + \gamma E_{s' \sim p(s,a)}[\max_{a' \in A} Q(s',a')]$$

$$\Leftrightarrow \quad Q = Q^*$$

we proposed two iteration algorithms for MDPs:

① Value iteration:

(i) initialize Q_0

(ii) apply $Q_k = \mathcal{T} Q_{k-1}$, $k = 1, 2, \ldots$

② Policy iteration:

(i) compute Q^{π_k}

(ii) update $\pi_{k+1} = \pi_{Q^{\pi_k}}$

we also proposed two LP approaches:

① Primal LP

$$\min \sum_s \mu(s) V(s)$$

$$\text{s.t.} \quad V(s) \geq r(s,a) + \gamma \sum_{s'} P(s'|s,a) V(s') \quad \forall (s,a)$$

② Dual LP

Let $K_\mu := \{d \mid d \geq 0,\ \sum_a d(s,a) = (1-\gamma)\mu(s) + \gamma \sum_{s',a'} P(s|s',a')d(s',a')\}$

$$\max \ \frac{1}{1-\gamma} \sum_{s,a} d(s,a)r(s,a)$$

$$\text{s.t.} \quad d \in K_\mu$$

Some analysis of MDPs:

① Although we provided some algorithms for optimization, when the $|S|$ or $|A|$ is infinite or continuous, it is hard to compute the iteration for machine.

② To fix ①, we can sample a fixed number of state-action pairs (s,a) and do the iteration on them.

③ Due to the same problem to ①, to find the very action a s.t. $a = \arg\max_{a' \in A} Q^l(s',a')$, we still need to sample many a'.

We name such sample processes as "A agent explore the environment"

In the end, we proved a significant equation describing the difference between two stationary policies:

① [Lemma 1.16] The performance difference lemma:

$$V^\pi(\mu) - V^{\pi'}(\mu) = \frac{1}{1-\gamma} E_{s' \sim d_\mu^\pi} E_{a' \sim \pi(s')}[A^{\pi'}(s',a')]$$