CS 598 NJ, Homework for 1st week

Nan Jiang

September 2, 2020

The purpose of this homework set is to help you digest course material. No need to submit.

1 Shift of rewards

Consider two MDPs $M = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$ and $M' = (\mathcal{S}, \mathcal{A}, P, R', \gamma)$, which only differ in their reward functions. Moreover, we have for any $s \in \mathcal{S}, a \in \mathcal{A}$,

$$R(s, a) = R'(s, a) + c,$$

where c is a universal constant that does not depend on s or a. For any policy π, let V_M^π denote its value function in M and $V_{M'}^\pi$ denote its value function in M'. For any $s \in \mathcal{S}$, can you express $V_M^\pi(s)$ using c and $V_{M'}^\pi(s)$?

After proving your result, think about its implications. In the lecture we made the assumption that rewards lie in $[0, R_{\max}]$. Why is this without loss of generality? What if I have an MDP whose rewards lie in $[-R_{\max}, R_{\max}]$?

2 Finite-horizon MDPs

In the lecture we considered infinite-horizon discounted MDPs: we sum up infinitely many rewards and a discount factor less than 1 keeps the sum finite. Now consider an alternative formulation where we cut down the trajectory after H steps, where H is a pre-defined constant. That is, with the same generative process of trajectories, we now consider return to be defined as

$$\mathbb{E}\left[\sum_{h=1}^{H} r_h\right].$$

A finite-horizon MDP is usually specified as $M = (\mathcal{S}, \mathcal{A}, P, R, H, d_0)$, where H is the episode length (or horizon) and $d_0 \in \Delta(\mathcal{S})$ is the initial state distribution (from which s_1 is drawn. Optimal policies in finite-horizon MDPs are generally *non-stationary*, i.e., you need to look at both the current state and the number of steps remaining to make an optimal decision.

State and prove the analogy of **Q1** for finite-horizon MDPs.

3 Indefinite-horizon MDPs

3.1

Here is yet another formulation, which is similar to finite-horizon MDPs except that the episode length H can vary: A subset of the state space $\mathcal{S}_{\text{term}} \subset \mathcal{S}$ are considered terminal, and an episode $s_1, a_1, r_1, s_2, a_2, r_2, \ldots$ keeps rolling out until we first visit a terminal state, $s_H \in \mathcal{S}_{\text{term}}$. In general, the length of the epsiode, H, is a random variable. The value is still defined as $\mathbb{E}[\sum_{h=1}^{H} r_h]$. Examples include the stochastic shortest paths shown in the slides. Is the analogy of the results in **Q1** and **Q2** still true?

As an example, consider a navigation task where the goal is to get to the destination state as soon as possible. Let's model it as an indefinite-horizon MDP: reward is -1 per step, and the process terminates whenever we reach the destination. It is clear then the return of a policy is the negative expected total number of steps towards destination. Makes sense.

Consider what happens when we add $+1$ to all rewards. What about $+2$?

3.2

Suppose there exists some constant H_0 such that $H \leq H_0$ holds almost surely for an indefinite-horizon MDP. Can you convert an indefinite-horizon MDP into an equivalent finite-horizon MDP? Hint: add an "absorbing" state which gives 0 reward and loops in itself.

Convert the navigation task in 3.1 into a finite-horizon MDP. What happens when we add $+1$ to all rewards in the corresponding finite-horizon MDP? What about $+2$? From **Q2** we know that these shifts should be valid. What's different from the situation in 3.1?

4 Non-stationary dynamics

So far all our definitions consider stationary dynamics, that is, the transition function only depends on the state and action, and does not depend on the time step. A finite-horizon MDP with non-stationary dynamics (and reward function) is a generalization: $M = (\mathcal{S}, \mathcal{A}, \{P_h\}_{h=1}^{H}, \{R_h\}_{h=1}^{H}, H, d_0)$, where $s_1 \sim d_0$, $s_{h+1} \sim P_h(s_h, a_h)$, and $r_{h+1} = R_h(s_h, a_h)$. That is, the transition rule and reward function can change as time elapses.

Answer the following questions:

(1) Why is this a generalization of stationary dynamics?

(2) Can you convert a non-stationary MDP into a stationary one? You may need to augment the state representation. How large is the state space after conversion?

(3) (Open) Does it make sense to define non-stationary dynamics for infinite-horizon, discounted MDPs?