CS 295: Optimal Control and Reinforcement Learning

Winter 2020

Assignment 1

due Thursday, January 23 2020, 11pm

Part 1 Relations between horizon settings:

1. Let $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, p, r \rangle$ be an MDP, and π: $\mathcal{S} \to \Delta(\mathcal{A})$ (where $\Delta(\mathcal{A})$ is the space of distributions over $\mathcal{A}$) a stochastic policy that induces the Markov process

$$p_\pi(s'|s) = \sum_{a \in A} \pi(a|s) p(s'|s,a) \qquad \forall s, s' \in \mathcal{S}.$$

   This question is concerned with a *stationary distribution* $p(s)$ of the process, which is a "fixed point" of the transition p_π, in the sense that if $p(s_t)$ is that stationary distribution then so is $p(s_{t+1})$.

   Background that goes a tiny bit deeper (but is not needed for the question): For a finite state space, think of p_π as a matrix P such that $P_{s,s'} = p_\pi(s'|s)$. If we have a probability (row) vector (non-negative and summing to 1) p_t that gives the distribution of s_t, then you can convince yourself that $p_{t+1} = p_t P$. A stationary distribution is then, by definition, a left-eigenvector of P with eigenvalue 1. You can convince yourself that 1 is always an eigenvalue of P, since $\vec{1}$ is always a *right*-eigenvector of P with eigenvalue 1 (left- and right-eigenvectors are generally different, but the set of eigenvalues from left and right is the same). So a stationary distribution always exists for finite state spaces (and also for infinite state spaces under mild conditions). The original question specified the case where the process has a unique stationary distribution, but that's not needed for the question.

   Now back to the question. Let's assume that the initial distribution $p(s_0)$ is a stationary distribution. Show that the expected return of the policy π is the same, up to some multiplicative constant, for a T-step finite horizon, an infinite horizon, and a discounted horizon.

   In this sense, what is the "effective finite horizon" of the discounted horizon with discount γ? (That's the finite horizon T that would gives the same multiplicative constant.)

2. In an episodic horizon, there exists an *absorbing state* $s_f \in \mathcal{S}$. A state is absorbing if any action $a \in \mathcal{A}$ taken in it remains in it $p(s_f|s_f,a) = 1$, and gets no reward

$r(s_f, a) = 0$. So once the trajectory gets to an absorbing state, nothing interesting can ever happen again, and we terminate the process. This allows us to define the return in the episodic horizon as $R = \sum_t r(s_t, a_t)$, although there's generally no bound on how many terms the series has before reaching $s_T = s_f$. We will restrict our attention to processes and policies that get to the absorbing state in finite *expected* time, so that the expected return is finite.

What is the stationary distribution with the policy thus restricted?

3. Show that the discounted horizon is a special case of the episodic horizon. No need to prove anything, just take an MDP $\mathcal{M}_1$ with a discounted horizon, and construct another MDP $\mathcal{M}_2$ with an episodic horizon such that the two problems are quite obviously equivalent (i.e. finding good policies in one is equivalent to finding good policies in the other). What state do you need to add to $\mathcal{M}_1$? What are the rewards and transition probabilities involving that state in $\mathcal{M}_2$, such that the discount is emulated?

Part 2 The lecture on Imitation Learning extended the one from the Berkeley Deep RL course (`https://rail.eecs.berkeley.edu/deeprlcourse/`). Solve both sections of Assignment 1 of that course: `https://rail.eecs.berkeley.edu/deeprlcourse/static/homeworks/hw1.pdf`.