Learning in Complex Systems (049004)

Homework 1: Dynamic Programming (Finite Horizon)

Submission date: March 21

1. Reading: Read Chapter 3 (Examples) in Puterman's book.

2-4. Solve Problems 3.2, 3.18, 3.20 from that chapter.

3. a. Read the proof of Theorem 2(i). Explain in a few sentences the main ideas.

   b. Complete the proof of Theorem 2(ii).

4. Value function for a general policy: Consider a general control policy, which may be history-dependent and random (i.e., a_t is selected according to a probability distribution $\pi_t = \pi_t(a \mid h_t)$). Generalize Lemma 1 on page 2.8 (and its proof) to this case (use a history-dependent value function).

5. Consider a stationary MDP with two states and two action, and finite time horizon N. Choose non-trivial (non-zero and unequal) transition probabilities and rewards. Draw a state transition diagram for your model, write down explicitly the value iteration equation for this model, and compute the optimal value function and optimal policy for $N = 3$ (assuming zero terminal rewards).

6. Consider an MDP with the following exponential reward functional:

$$J = E^{\pi,s}\left\{\exp\beta\left(\sum_{t=0}^{N-1} r_k(s_k, a_k) + r_N(s_N)\right)\right\}$$

   This reward functional is called *risk averse* or *risk seeking* (depending on the sign of β), as it assigns higher or lower probabilistic weight to low (negative) outcomes. .

   a. What would the optimal policy converge to as $\beta \to 0$ (hint: use a Taylor expansion).

   b. Suggest a recursive programming algorithm that obtains the optimal value function and the optimal policy for this problem. Also express the recursion in terms of $v_t(s) = \log V_t(s)$, and compare to the standard case.