Homework 3

*Collaboration in the sense of discussion is allowed, however, the assignment is **individual** and the work you turn in should be entirely your own. See the collaboration and academic integrity statement here:* https://natanaso.github.io/ece276b. *Books may be consulted but not copied from. Please acknowledge **in writing** people you discuss the problems with and provide references for any books or papers you use.*

Submission

Upload your solutions on **Gradescope** by the deadline shown at the top right corner. You may use latex, scanned handwritten notes (write legibly!), or any other method to prepare a pdf file. Do not just write the final result. Present your work in detail, explaining your approach at every step.

Problems

In square brackets are the points assigned to each part.

1. Consider a discrete-time system with state $x_t \in \mathbb{R}$, input $u_t \in \mathbb{R}$, and motion model:

$$x_{t+1} = x_t + u_t. \tag{1}$$

We are interested in designing a control policy $\pi : \mathbb{R} \to \mathbb{R}$ for an infinite-horizon discount optimal control problem with stage cost:

$$\ell(x, u) = \sin^2(x) + u^2 \tag{2}$$

and discount factor $\gamma = 0.5$.

   (a) [10 pts] Consider the policy $\pi(x) = -x/2$. Determine the action-value function $Q^\pi(x, u)$. You may assume that $\pi(x)$ performs well so that after one time step the state is small and we may use a small-angle approximation $\sin(x) \approx x$. However, the approximation should not be used at the very first state.

   (b) [10 pts] Using $Q^\pi(x, u)$ you computed in part (a), perform one step of policy improvement to obtain a new policy $\pi'(x)$. Define $\pi'(x)$ precisely.

2. Consider a discrete-time system with state $x_t \in \mathbb{R}$, input $u_t \in \mathbb{R}$, and motion model:

$$x_{t+1} = \sqrt{2}x_t + u_t + 2w_t, \tag{3}$$

where w_t is Gaussian motion noise with zero mean and variance $1/2$. We are interested in solving an infinite-horizon discounted optimal control problem with discount factor $\gamma = 1/2$ and stage cost:

$$\ell(x, u) = \frac{1}{2}x^2 + \frac{1}{2}u^2. \tag{4}$$

Assume that the optimal value function is of the form $V^*(x) = ax^2 + bx + c$.

   (a) [20 pts] Determine the parameters a, b, c.

   (b) [10 pts] Determine the optimal Q function $Q^*(x, u)$.

3. Consider a Markov Decision Process with state space $\mathcal{X} = \{1, 2\}$ and control space $\mathcal{U} = \{a, b\}$. Let the transition probability matrices be:

$$\mathbf{P}^a := \begin{bmatrix} 1/8 & 7/8 \\ 5/8 & 3/8 \end{bmatrix}, \qquad \mathbf{P}^b := \begin{bmatrix} 3/8 & 5/8 \\ 5/8 & 3/8 \end{bmatrix}, \tag{5}$$

where $\mathbf{P}^u_{i,j}$ specifies the probability of transitioning from state $i \in \mathcal{X}$ to state $j \in \mathcal{X}$ under control $u \in \mathcal{U}$. Consider an infinite-horizon discounted stochastic optimal control problem with discount factor $\gamma = 0.8$ and stage cost:

$$\ell(x, u) := \begin{cases} 16x & u = a, \\ 5x & u = b. \end{cases} \tag{6}$$

   (a) [10 pts] Formulate an equivalent first-exit infinite-horizon optimal control problem. Provide all elements necessary to define the problem, including the state space, control space, motion model, stage cost, terminal cost, and terminal states.

   (b) [10 pts] Starting with an initial value function estimate $V_0(1) = 20$, $V_0(2) = 10$ for the first-exit problem in part (a), apply one iteration of the Value Iteration algorithm to compute $V_1(x)$.

   (c) [20 pts] Formulate a linear program whose optimal solution is the optimal value function $V^*(x)$. Solve the linear program using `cvxpy`[1] to obtain $V^*(x)$.

4. [10 pts] Consider a first-exit infinite-horizon stochastic optimal control problem with state space $\mathcal{X} = \{1, 2, 3\}$. The motion model and cost functions are unknown but you have observed two episodes from a policy π:

$$1 \overset{-3}{\to} 1 \overset{-2}{\to} 2 \overset{+4}{\to} 1 \overset{-3}{\to} 2 \overset{+3}{\to} 3,$$
$$2 \overset{+4}{\to} 1 \overset{-2}{\to} 2 \overset{+3}{\to} 3,$$

where the arrows indicate a transition and the costs are shown above. For example, $1 \overset{-3}{\to} 1$ indicates a transition from state 1 to state 1 with a cost of -3. Estimate the value function $V^\pi(x)$ of policy π using Temporal Difference Policy Evaluation (TD(0)) with step size $\alpha = 0.2$ and initial estimate $V(1) = V(2) = V(3) = 1/2$.

[1] https://www.cvxpy.org/