CSC 696H Homework 1

Chicheng Zhang

September 2021

- This homework is due on Sep 30 in class.
- Your solutions to these problems will be graded based on both correctness and clarity. Your arguments should be clear: there should be no room for interpretation about what you are writing. Otherwise, I will assume that they are wrong, and grade accordingly.
- If you feel unable to make progress on any of the questions, you can post your questions on Piazza. Try posing your questions to be as general as possible, so that it can promote discussion among the class.
- You are encouraged to discuss the homework questions with your classmates, but the discussions should only be at a high level, and you should write your solutions in your own words. For every question you have had discussions on, please mention explicitly whom you have discussed with; otherwise it may be counted as academic integrity violation.
- Feel free to use existing theorems from the course notes / the textbook.

Problem 1

Suppose we are given a discounted MDP $M = (\mathcal{S}, \mathcal{A}, \gamma, r, \mathbb{P}, \mu)$ and a stationary policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$.

1. Use the Bellman consistency equation given in class (the two-way relationship between Q^π and V^π) to derive the *Bellman consistency equation for* V^π:

$$\forall s \in \mathcal{S}, V^\pi(s) = r^\pi(s) + \gamma \sum_{s' \in \mathcal{S}} M^\pi(s' \mid s) V^\pi(s'),$$

   where $r^\pi(s) = \sum_{a \in \mathcal{A}} \pi(a \mid s) r(s, a)$, and $M^\pi(s' \mid s) = \sum_{a \in \mathcal{A}} \pi(a \mid s) \mathbb{P}(s' \mid s, a)$.

2. Suppose $\mathcal{S} = \{1, 2, 3\}$, $\mathcal{A} = \{L, R\}$, $\gamma = 0.9$, and the reward and the transition probability are given in Tables 1 and 2. Define π as: $\pi(a \mid s) = 0.5$ for all s, a. Calculate V^π (you can use tools such as matlab or numpy).

(s, a)	$r(s, a)$
$(1, L)$	0.1
$(1, R)$	0
$(2, L)$	0
$(2, R)$	0
$(3, L)$	0
$(3, R)$	1

Table 1: Reward function $r(s, a)$.

(s, a) \ s'	1	2	3
$(1, L)$	1	0	0
$(1, R)$	0.4	0.6	0
$(2, L)$	1	0	0
$(2, R)$	0.05	0.6	0.35
$(3, L)$	0	1	0
$(3, R)$	0	0.4	0.6

Table 2: Transition probability $\mathbb{P}(s' \mid s, a)$.

Problem 2

Suppose we are given a discounted MDP $M = (\mathcal{S}, \mathcal{A}, \gamma, r, \mathbb{P}, \mu)$ and a deterministic stationary policy $\pi : \mathcal{S} \to \mathcal{A}$, such that for all $s \in \mathcal{S}$, $\pi(s) \in \operatorname{argmax}_{a \in \mathcal{A}} Q^\pi(s, a)$ (i.e. it is greedy with respect to its own action-value function). Is π optimal in M? Justify your answer.

Problem 3

Given a discounted MDP $M = (\mathcal{S}, \mathcal{A}, \gamma, r, \mathbb{P}, \mu)$ and a stationary policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$. Define operator $\mathcal{T}^\pi$ such that for all $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, $\mathcal{T}^\pi f$ is a mapping from $\mathcal{S} \times \mathcal{A}$ to $\mathbb{R}$, such that for all (s, a),

$$(\mathcal{T}^\pi f)(s, a) = r(s, a) + \gamma \sum_{s' \in \mathcal{S}} \mathbb{P}(s' \mid s, a) \sum_{a' \in \mathcal{A}} \pi(a' \mid s') f(s', a').$$

1. Prove that $\mathcal{T}^\pi$ is a contraction, specifically, given any two functions $f, g : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, $\|\mathcal{T}^\pi f - \mathcal{T}^\pi g\|_\infty \leq \gamma \|f - g\|_\infty$.
2. Suppose $f_1 = 0$ and $f_{i+1} = \mathcal{T}^\pi f_i$ for all $i \in \mathbb{N}$. Does the sequence $\{f_i\}_{i=1}^\infty$ converge? If so, how fast is the convergence?
3. Suppose we are given two stationary policies π_1, π_2. Prove that

   $$(\mathcal{T}^{\pi_1} Q^{\pi_2})(s, a) = \mathbb{E}\left[G_0 \mid (s_0, a_0) = (s, a), a_1 \sim \pi_1, a_{2:\infty} \sim \pi_2\right],$$

   i.e., it is the expected return of the nonstationary policy that executes π_1 at time step 1, and executes π_2 afterwards, conditioned on being at state s and taking action a at time step 0.
4. Suppose we are given three stationary policies π_1, π_2, π_3. Describe $(\mathcal{T}^{\pi_3}(\mathcal{T}^{\pi_2} Q^{\pi_1}))(s, a)$ in words (similar to the verbal description in the previous item), and briefly justify.

Problem 4

In the lectures that established the Bellman optimality equation, we have shown that

$$\begin{cases} V^*(s) \leq \max_{a \in \mathcal{A}} Q^*(s, a), & \forall s \\ Q^*(s, a) \leq r(s, a) + \gamma \sum_{s' \in \mathcal{S}} \mathbb{P}(s' \mid s, a) V^*(s), & \forall s, a \\ \pi^*(s) = \operatorname{argmax}_{a \in \mathcal{A}} Q^*(s, a), & \forall s \end{cases}$$

implies $Q^* \preceq Q^{\pi^*}$. Using a similar reasoning, show that $V^* \preceq V^{\pi^*}$ (for this question, you should pretend that you haven't learned the Bellman optimality equation yet - imagine we are trying to derive the Bellman optimality equation from scratch).

Problem 5

1. Consider two undiscounted episodic MDPs M and $\hat{M}$ that share the same state space $\mathcal{S}$, action space $\mathcal{A}$, episode length H, initial state distribution μ, but different reward functions and transition probabilities; specifically, M has $(r_h, \mathbb{P}_h)_{h=0}^{H-1}$ while $\hat{M}$ has $(\hat{r}_h, \hat{\mathbb{P}}_h)_{h=0}^{H-1}$. Generalize the simulation lemma given in the class to give an identity for

   $$V_h^\pi(s) - \hat{V}_h^\pi(s)$$

   in terms of $r_h - \hat{r}_h$ and $\mathbb{P}_h - \hat{\mathbb{P}}_h$, and prove its correctness.

2. Now consider a variant of the "RL with a generative model" setting, where the simulator takes into input (s, a, h), and returns $s' \sim \mathbb{P}_h(\cdot \mid s, a)$ along with a *noisy reward* $r' \sim \text{Bernoulli}(r_h(s, a))$. Modify the sample-based VI algorithm given in the class, so that it can use this new simulator to learn a ϵ-optimal policy $\hat{\pi}$ with probability $1 - \delta$. Present your algorithm and justify your answer. What is the your algorithm's total number of calls to the simulator?