The Johns Hopkins University
Department of Electrical and Computer Engineering

520.637 — Foundations of Reinforcement Learning — Fall 2023

Midterm Exam

Name: Ha Manh Bui

Issued. Friday, October 27, 2022, 9:00 am EDT
Due. Friday, October 27, 2022, 11:59pm EDT

Instructions (Read carefully!)

1. This exam is open book and class notes.
2. No online sources besides the ones provided in blackboard are allowed.
3. Where proofs are required, give neat, step-by-step solutions that include all details and are based on first principles.
4. Where computations are required, give neat, step-by-step solutions that include all the details.
5. You **may NOT discuss the problems with anyone.**
6. Upload a unique pdf file that includes a signed copy of this cover page. **Good Luck!**

Marks

Question	Maximum	Marks
1	20	
2	20	
3	20	
4	20	
Total	80	

Statement of Academic Honesty: For this exam, I make the following truthful statements:

1. I have not received, I have not given, nor will I give or receive, any assistance to another student taking this exam.
2. I will not use any non-instructor approved sources to assist me on an exam.
3. I will not plagiarize someone else's work and turn it in as my own.
4. I understand that acts of academic dishonesty may be penalized to the full extent allowed by the Johns Hopkins University Student Conduct Code, including receiving a failing grade for the course. I recognize that I am responsible for understanding the provisions of the Johns Hopkins University Student Conduct Code as they relate to this academic exercise.

Place your initials here: HMB

Problem 1.

Consider an infinite horizon discounted problem ($0 < \gamma < 1$) in which the state space is finite, with n states, and there are only two possible decisions: stop or continue.

- If at time t you are at state s, and you decide to stop, you incur a stopping cost $g(s)$ and move to a cost-free state, at which you stay forever.
- If at time t you are at state s, and you decide to continue, you incur a continuation cost $c(s)$ and move to a next state s', selected at random according to transition probabilities $p(s, s') = \mathbb{P}(S_{t+1} = s' | S_t = s)$.

Assume that all $g(s)$ and $c(s)$ are nonnegative. We use policy iteration to find the optimal policy that minimizes the discounted total cost $\sum_{t=1}^{\infty} \gamma^t C_t$ for this MDP.
Note that a stationary policy π is completely specified in terms of the set S_π of states at which the policy decides to stop. Let $\pi_0, \pi_1, \cdots$ be the sequence of policies generated by the policy iteration algorithm, starting with a given arbitrary policy π_0. Suppose that any ties are broken in favor of the "continue" decision when a greedy policy is constructed in the policy improvement phase.

1. Given any value function v, what is its greedy policy?

2. Given a policy π, what is its value function v_π? Find a closed form.
Hint: consider (1) the states at which π stops and (2) the rest, write down their value function separately.

3. Is it always true that for any initial policy π_0, we have $S_{\pi_1} \subseteq S_{\pi_0}$? Give the proof or a counterexample.

4. Suppose $S_{\pi_0} = \mathcal{S}$, i.e. the policy π_0 stops at every state. It is obvious that $S_{\pi_1} \subseteq S_{\pi_0}$. Now show that

$$S_{\pi_2} \subseteq S_{\pi_1} \subseteq S_{\pi_0} .$$

You may use the fact that for any non-negative matrix M with $\sum_j M_{ij} \leq 1$, $(I - \gamma M)^{-1}$ is also non-negative.

Problem 2.

Given a MDP $\mathcal{M}(\mathcal{S}, \mathcal{A}, p, r)$, we assume that the reward function $r(s,a)$ is known, deterministic and bounded $|r(s,a)| \leq 1$.
Suppose we have access to a generative model, which can provide us with a sample $s' \sim p(\cdot|s,a)$ upon input of any state action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$. Suppose we call our simulator N times at each state action pair. Let $\hat{p}$ be our empirical model, defined as follows:

$$\hat{p}(s'|s,a) = \frac{\text{count}(s',s,a)}{N}, \ \forall s', (s,a)$$

where $\text{count}(s',s,a)$ is the number of times the state-action pair (s,a) transitions to state s'. As the N is the number of calls for each state action pair, the total number of calls to our generative model is $|\mathcal{S}||\mathcal{A}|N$. We define $\hat{\mathcal{M}}$ to be the empirical MDP that is identical to the original $\mathcal{M}$, except that it uses $\hat{p}$ instead of p for the transition model. We let $\hat{V}^\pi$ denote the value function of policy π under $\hat{\mathcal{M}}$.

1. Show that for any policy π, we have

$$\|V^\pi\|_\infty \leq \frac{1}{1-\gamma}\,.$$

2. Consider a fixed policy π,

(a) Define

$$\|P_\pi - \hat{P}_\pi\|_1 := \max_s \sum_{s'} \left| \sum_a p(s'|s,a)\pi(a|s) - \sum_a \hat{p}(s'|s,a)\pi(a|s) \right|,$$

show that

$$\|P_\pi - \hat{P}_\pi\|_1 \leq \max_s \max_a \sum_{s'} |p(s'|s,a) - \hat{p}(s'|s,a)|\,.$$

(b) It can be shown that for any fixed (s,a)-pair, we have

$$\mathbb{P}\left(\sum_{s'} |p(s'|s,a) - \hat{p}(s,a)| \geq \epsilon \right) \leq \exp\left(-N \frac{\epsilon^2}{4|\mathcal{S}|} \right),$$

for any $N \geq \frac{4|\mathcal{S}|}{\epsilon^2}$.
Use this to show: Given $\delta \in (0,1)$ and $\epsilon > 0$. If

$$N \geq \frac{4|\mathcal{S}|}{\epsilon^2} \log \frac{|\mathcal{S}||\mathcal{A}|}{\delta}\,,$$

then with probability $1-\delta$, we have

$$\|P_\pi - \hat{P}_\pi\|_1 \leq \epsilon\,.$$

3. Show that for a fixed policy π,

$$\|V^\pi - \hat{V}^\pi\|_\infty \leq \frac{\gamma}{1-\gamma} \|(P_\pi - \hat{P}_\pi)V^\pi\|_\infty$$

4. Finally, show: Given $\delta \in (0,1)$ and $\epsilon > 0$. If

$$N \geq \frac{4\gamma^2|\mathcal{S}|}{(1-\gamma)^4\epsilon^2} \log\left(\frac{|\mathcal{S}||\mathcal{A}|}{\delta} \right),$$

then with at least probability $1-\delta$, we have

$$\|V^\pi - \hat{V}^\pi\|_\infty \leq \epsilon\,.$$

Problem 3.

For the multi-arms bandit problem we discussed in the class. Suppose that we get return G_n at n-th time we do action a, and we follow the usual assumptions: 1) G_n's are independent; 2) $\mathbb{E}G_n = r,\ n = 1, 2, \cdots$. Let Q_{n+1} be our estimates of r after we do action a the n-th time, and we have the following update rule

$$Q_{n+1} = Q_n + \alpha_n(G_n - Q_n)\,,\ Q_1 = 0\,. \tag{1}$$

We would like to show that, in this case, a weaker condition than $\sum_n \alpha_n^2 < +\infty$ exists for convergence in expectation.

1. Show that

$$\prod_{k=2}^{n}\left(1 - \frac{1}{\sqrt{k}}\right) \le \frac{1}{\sqrt{n}}\,,$$

(Hint: $\sqrt{k} - 1 \le \sqrt{k-1}$ when $k \ge 1$)

2. Show that for any $c > 0$, there exists $k_0 > 0$ that depends on c, such that

$$\sqrt{k} - c \le \sqrt{k-1}, \quad \forall k > k_0\,.$$

3. Show that for any $c > 0$,

$$\lim_{n\to\infty} \prod_{k=1}^{n}\left|1 - \frac{c}{\sqrt{k}}\right| = 0\,.$$

4. Suppose that we choose in (1) $\alpha_n = \frac{c}{\sqrt{n}}$ for some $c > 0$, show that

$$\lim_{n\to\infty} |\mathbb{E}Q_n - r| = 0\,.$$

Problem 4. Suppose you have s_0 items that you would like to sell ($s_0 \in \mathbb{N}$). You can sell at most one item per day, and you have a total of N days in which you can sell them.
At the $k-$th day, given the remaining s_k items you have for sale, you can decide the unit price as $a_k \in \mathbb{R}_{\geq 0}$. After deciding the price, the probability of an item being sold is given by:

$$\lambda(a_k) = \beta e^{-a_k} \tag{2}$$

where $0 < \beta \leq 1$. The goal in this setting is to find the optimal way to set prices, with the goal of maximizing the expected total reward.

1. What is the MDP associated with this problem? Specify the horizon, state space, action space and state-reward transition kernel.

What is the objective function you are trying to maximize?

2. Let $V_k(s_k)$ be the optimal value function for the $k-$th stage at state s_k. Write the Dynamic Programming equations for the optimal policy for all stages.

3. Assume the value functions $V_k(s_k)$ are monotonically nondecreasing on s_k. Show that for all $s_k > 0$, the optimal price-setting policy satisfies:

$$\pi_k^\star(s_k) = 1 + V_{k+1}(s_k) - V_{k+1}(s_k - 1) \tag{3}$$

Furthermore, show that:

$$V_k(s_k) = \beta e^{-\pi^\star(s_k)} + V_{k+1}(s_k)\ . \tag{4}$$

State clearly how you are using the given assumption.

4. Show that, for all k:

(a) $V_k(s_k)$ is indeed non-decreasing in s_k.

(b) The optimal policy $\pi_k^\star(s_k)$ is non-increasing in s_k.

(c) The closed-form solution for V_k is:

$$V_k(s_k) = \begin{cases} (N-k)\beta e^{-1} & \text{if } s_k \geq N-k, \\ \sum_{i=k}^{N-s_k} \beta e^{-\pi_i^\star(s_k)} + s_k \beta e^{-1} & \text{if } 0 < s_k < N-k, \\ 0 & \text{if } s_k = 0 \end{cases}$$

Hint: Use induction to simultaneously show (a) and (b).

Problem 1:

1, Under the infinite horizon discounted setting, by definition, we have two possible action "stop" or "continue", so the Q-function follows:

$$\begin{cases} q_\pi(S=s, A="\text{stop}") = g(s) \\ q_\pi(S=s, A="\text{continue}") = c(s) + \gamma \sum_{s'} p(s'|s)\, v(s') \end{cases}$$

$\Rightarrow \forall v$, the greedy policy π_v, $\pi_v(a|s) \geq 0$ is

$$\begin{cases} \text{"stop"} & \text{if } g(s) < c + \gamma \sum_{s'} p(s'|s)\, v(s') \\ \text{"continue"} & \text{otw} \end{cases}$$

2, If $s \in S_\pi$, then $v_\pi(s) = g(s)$. If $s \notin S_\pi$, we have

$$v_\pi(s) = c(s) + \gamma \sum_{s' \notin S_\pi} p(s'|s)\, v_\pi(s) + \gamma \sum_{s' \in S_\pi} p(s'|s)\, g(s') \quad (1)$$

Let $P'(s'|s)$ be a $m \times m$ transition matrix for $s' \in S_\pi$, where $m = |S| - |S_\pi|$ & $r_\pi(s) = c(s) + \gamma \sum_{s' \in S_\pi} p(s'|s)\, g(s')$, then from (1), we get

$$v_\pi(s) = r_\pi(s) + \gamma \sum_{s' \in S_\pi} P'(s'|s)\, v_\pi(s)$$

$\Rightarrow$ the closed form of $v_\pi(s)$ is

$$v_\pi = (I - \gamma P')^{-1} r_\pi. \quad (2)$$

3, It is not always true that for any initial policy π_0, we have $S_{\pi_1} \subseteq S_{\pi_0}$.
One counterexample is:
Let $S_{\pi_0} = \{\phi\}$, the policy π_0 decide to continue at any state.
Assume $c(s) = 1$

$$\Rightarrow v_{\pi_0}(s) = q_{\pi_0}(s, \text{"continue"}) = \sum_{t=1}^{\infty} \gamma^t c_t = \frac{1}{1-\gamma}.$$

If $g(s_1) < \frac{1}{1-\gamma}$, after the policy improvement step, π_1 will stop at state s_1

$$\Rightarrow S_{\pi_1} \supset S_{\pi_0} \text{ (contradict!)}.$$

4, Consider at the beginning: $v_{\pi_0}(s) = g(s)$, $\forall s \in S$.
Assume $\pi_1(s_1) =$ "continue"

$$\Rightarrow g(s_1) = q_{\pi_0}(s_1, \text{"stop"}) \geq q_{\pi_0}(s_1, \text{"continue"}).$$

Using the fact that $\forall$ non-negative matrix P' with $\sum_j P'_{ij} \leq 1$, $(I - \gamma P')^{-1}$ is also non-negative. Apply for (2), we get

$$v_{\pi_0}(s_1) \geq v_{\pi_1}(s_1) \geq v_{\pi_2}(s_1).$$

$$\Rightarrow q_{\pi_1}(s_1, \text{"continue"}) \geq q_{\pi_2}(s_1, \text{"continue"})$$

$$\Rightarrow g(s_1) \geq q_{\pi_1}(s_1, \text{"continue"}) \geq q_{\pi_2}(s_1, \text{"continue"}),$$

i.e, when π_1 change from "stop" to "continue", π_2 can not change from "continue" to "stop". Combining with the fact that if π_1 doesn't "continue", then $S_{\pi_1} = S$ & $\pi_1 = \pi_0 \Rightarrow S_{\pi_2} \subseteq S_{\pi_1}$. As a result, we achieve

$$S_{\pi_2} \subseteq S_{\pi_1} \subseteq S_{\pi_0}. \quad \square$$

Problem 2:

1, Under the infinite horizon discounted setting, by definition of the value function, $\forall s \in \mathcal{S}$, we have:

$$V^{\pi}(s) = \mathbb{E}_{\pi}\left\{ \sum_{t=0}^{\infty} \gamma^t R_t \mid S=s \right\}.$$

Since $|r(s,a)| \leq 1$, we get:

$$V^{\pi}(s) = \mathbb{E}_{\pi}\left\{ \sum_{t=0}^{\infty} \gamma^t R_t \mid S=s \right\} \leq \sum_{t=0}^{\infty} \gamma^t = \lim_{t\to\infty} \frac{1-\gamma^t}{1-\gamma} = \frac{1}{1-\gamma}, \quad \forall s \in \mathcal{S}.$$

Combining with the fact that $\|X\|_{\infty} = \max_i \|X_i\|$, we obtain:

$$\|V^{\pi}\|_{\infty} \leq \frac{1}{1-\gamma} \quad \square \ (3)$$

2a, Apply triangle inequality, $\forall s \in \mathcal{S}$, we have:

$$\sum_{s'} \left| \sum_a p(s'|s,a)\pi(a|s) - \sum_a \hat{p}(s'|s,a)\pi(a|s) \right|$$

$$\leq \sum_{s'} \sum_a \left| p(s'|s,a)\pi(a|s) - \hat{p}(s'|s,a)\pi(a|s) \right|$$

$$= \sum_a \pi(a|s) \sum_{s'} \left| p(s'|s,a) - \hat{p}(s'|s,a) \right|$$

$$\leq \sum_a \pi(a|s) \left\{ \max_a \sum_{s'} \left| p(s'|s,a) - \hat{p}(s'|s,a) \right| \right\}$$

$$= \max_a \sum_{s'} \left| p(s'|s,a) - \hat{p}(s'|s,a) \right|$$

$$\Rightarrow \max_a \sum_{s'} \left| \sum_a p(s'|s,a)\pi(a|s) - \sum_a \hat{p}(s'|s,a)\pi(a|s) \right| = \|P_{\pi} - \hat{P}_{\pi}\|_1$$

$$\leq \max_s \max_a \sum_{s'} \left| p(s'|s,a) - \hat{p}(s'|s,a) \right| \quad \square \ (4)$$

2b, Given $\varepsilon > 0$, by definition δ from (5), we have

$$\|P_\pi - \hat{P}_\pi\|_1 \geq \varepsilon \Rightarrow \max_s \max_a \sum_{s'} |p(s'|s,a) - \hat{p}(s'|s,a)| \geq \varepsilon.$$

Using the fact that if $A \Rightarrow B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$

$$\Rightarrow \mathbb{P}(\|P_\pi - \hat{P}_\pi\|_1 \geq \varepsilon) \leq \mathbb{P}\left(\max_s \max_a \sum_{s'} |p(s'|s,a) - \hat{p}(s'|s,a)| \geq \varepsilon\right) \quad (5)$$

On the other hand, $\forall$ fixed (s,a)-pair s.t. $\sum_{s'} |p(s'|s,a) - \hat{p}(s'|s,a)| \geq \varepsilon$, then $\max_s \max_a |p(s'|s,a) - \hat{p}(s'|s,a)| \geq \varepsilon$, yielding

$$\mathbb{P}\left(\exists s,a : \sum_{s'} |p(s'|s,a) - \hat{p}(s'|s,a)| \geq \varepsilon\right) = \mathbb{P}\left(\max_s \max_a |p(s'|s,a) - \hat{p}(s'|s,a)| \geq \varepsilon\right).$$

$\Rightarrow \forall N \geq \frac{4|S|}{\varepsilon^2}$, applying union bound, we get:

$$\mathbb{P}\left(\exists s,a : \sum_{s'} |p(s'|s,a) - \hat{p}(s'|s,a)| \geq \varepsilon\right) \leq \sum_{s,a} \mathbb{P}\left(\sum_{s'} p(s'|s,a) - \hat{p}(s'|s,a) \geq \varepsilon\right).$$

$$\leq \sum_{s,a} \exp\left(-N \frac{\varepsilon^2}{4|S|}\right) = |S||A| \exp\left(-N \frac{\varepsilon^2}{4|S|}\right) \leq \delta, \quad \forall \delta \in (0,1).$$

$$\Rightarrow \quad N \geq \frac{4|S|}{\varepsilon^2} \log \frac{\delta}{|S||A|}.$$

Combining with (5) $\Rightarrow \mathbb{P}(\|P_\pi - \hat{P}_\pi\|_1 \geq \varepsilon) \leq \delta$,

i.e., if $N \geq \frac{4|S|}{\varepsilon^2} \log \frac{|S||A|}{\delta}$, then with prob. $1-\delta$, we have

$$\|P_\pi - \hat{P}_\pi\|_1 \leq \varepsilon. \quad \square.$$

3, Using the Bellman Eq., we have

$$\|V^{\pi} - \hat{V}^{\pi}\|_{\infty} = \|(R + \gamma P_{\pi} V^{\pi}) - (R + \gamma \hat{P}_{\pi} \hat{V}^{\pi})\|_{\infty}$$

$$\leq \gamma \|P_{\pi} V^{\pi} - \hat{P}_{\pi} \hat{V}^{\pi}\|_{\infty}$$

$$= \gamma \|P_{\pi} V^{\pi} - \hat{P}_{\pi} V^{\pi} + \hat{P}_{\pi} V^{\pi} - \hat{P}_{\pi} \hat{V}^{\pi}\|_{\infty}$$

$$\leq \gamma \|(P_{\pi} - \hat{P}_{\pi}) V^{\pi}\|_{\infty} + \gamma \|\hat{P}_{\pi} (V^{\pi} - \hat{V}^{\pi})\|_{\infty}. \quad (6)$$

Since $\forall s \in S$, we have.

$$\sum_{s'} p(s'|s)\left(V^{\pi}(s') - \hat{V}^{\pi}(s')\right) \leq \sum_{s'} p(s'|s) \max_{s'} \left\{V^{\pi}(s') - \hat{V}^{\pi}(s')\right\}$$

$$= \max_{s'} \left\{V^{\pi}(s') - \hat{V}^{\pi}(s')\right\},$$

i.e., $\|\hat{P}_{\pi}(V^{\pi} - \hat{V}^{\pi})\|_{\infty} \leq \|V^{\pi} - \hat{V}^{\pi}\|_{\infty}$. Applying this result to (6),

we obtain:

$$\|V^{\pi} - \hat{V}^{\pi}\|_{\infty} \leq \gamma \|(P_{\pi} - \hat{P}_{\pi}) V^{\pi}\|_{\infty} + \gamma \|\hat{P}_{\pi}(V^{\pi} - \hat{V}^{\pi})\|_{\infty}$$

$$\leq \gamma \|(P_{\pi} - \hat{P}_{\pi}) V^{\pi}\|_{\infty} + \gamma \|V^{\pi} - \hat{V}^{\pi}\|_{\infty}$$

$$\leq \frac{\gamma}{1-\gamma} \|(P_{\pi} - \hat{P}_{\pi}) V^{\pi}\|_{\infty} \quad \square \quad (7)$$

4, Using the result in (3), we have:

$$\|V^\pi - \hat{V}^\pi\|_\infty \le \frac{\gamma}{1-\gamma} \|(P_\pi - \hat{P}_\pi) V^\pi\|_\infty$$
$$\le \frac{\gamma}{1-\gamma} \|P_\pi - \hat{P}_\pi\|_1 \|V^\pi\|_\infty$$
$$\le \frac{\gamma}{(1-\gamma)^2} \|P_\pi - \hat{P}_\pi\|_1 \quad \text{(from (2))}$$

$\Rightarrow$ If $\|V^\pi - \hat{V}^\pi\|_\infty \ge \varepsilon$, then $\|P_\pi - \hat{P}_\pi\|_1 \ge \frac{(1-\gamma)^2}{\gamma}\varepsilon$, this yields

$$P(\|V^\pi - \hat{V}^\pi\|_\infty \ge \varepsilon) \le P(\|P_\pi - \hat{P}_\pi\|_1 \ge \frac{(1-\gamma)^2}{\gamma}\varepsilon)$$
$$\le \delta, \quad \forall \delta \in (0,1),$$

i.e., $P(\|V^\pi - \hat{V}^\pi\|_\infty \le \varepsilon) \ge 1-\delta$. Applying the result in 2b, this inequality holds, i.e., with at least prob. $1-\delta$, we have $\|V^\pi - \hat{V}^\pi\|_\infty \le \varepsilon$ if $N \ge \frac{c\gamma^2|S|}{(1-\gamma)^5\varepsilon^2} \log\left(\frac{|S||A|}{\delta}\right)$.

□

Problem 3,

1/ We have:

$$\prod_{k=2}^{n}\left(1-\frac{1}{\sqrt{k}}\right)=\prod_{k=2}^{n}\left(\frac{\sqrt{k}-1}{\sqrt{k}}\right).$$

Using the fact that $\sqrt{k}-1 \le \sqrt{k-1}$ when $k \ge 1$

$$\Rightarrow \prod_{k=2}^{n}\left(1-\frac{1}{\sqrt{k}}\right) \le \prod_{k=2}^{n}\left(\frac{\sqrt{k-1}}{\sqrt{k}}\right)=\frac{1}{\sqrt{n}} \quad \square.$$

2/ $\forall c>0$, we have:

$$\sqrt{k}-c \le \sqrt{k-1}$$
$$\Leftrightarrow \sqrt{k} \le \sqrt{k-1}+c$$
$$\Leftrightarrow k \le k-1+2c\sqrt{k-1}+c^2$$
$$\Leftrightarrow 1-c^2 \le 2c\sqrt{k-1}.$$

If $1-c^2<0 \Rightarrow$ this inequality holds $\forall k \ge 1$ (condition for $\sqrt{k-1}$ is defined)

otw, i.e., $1-c^2 \ge 0$

$$\Rightarrow k \ge \frac{(1-c^2)^2}{4c^2}+1$$

$\Rightarrow \forall c>0$, there is always exists $k_0>0$ that depends on c,

s.t. $\sqrt{k}-c \le \sqrt{k-1}$, $\forall k \ge k_0$ $\square$.

3/ We have

$$\lim_{n\to\infty} \prod_{i=1}^{n} \left| 1 - \frac{c}{\sqrt{k}} \right| = \lim_{n\to\infty} \prod_{i=1}^{n} \left| \frac{\sqrt{k}-c}{\sqrt{k}} \right| = \lim_{n\to\infty} \prod_{k=1}^{k_0} \left| \frac{\sqrt{k}-c}{\sqrt{k}} \right| \cdot \prod_{k=k_0+1}^{n} \left| \frac{\sqrt{k}-c}{\sqrt{k}} \right|$$

$$= \lim_{n\to\infty} \prod_{k=1}^{k_0} \left| \frac{\sqrt{k}-c}{\sqrt{k}} \right| \frac{\sqrt{k_0+2}-c}{\sqrt{n}} \quad (\text{from } 3.1, 3.2)$$

$$= 0 \quad \left(\text{due to } \prod_{k=1}^{k_0} \left| \frac{\sqrt{k}-c}{\sqrt{k}} \right| < \infty \text{ when } \sqrt{k}-c \le \sqrt{k}\right)$$

4/ By definition & assumption ① & ②, we have

$$\mathbb{E}\{Q_{n+1}\} = \mathbb{E}\{Q_n + \alpha_n (G_n - Q_n)\} = \mathbb{E}\{Q_n\} + \alpha_n (\mathbb{E}\{G_n\} - \mathbb{E}\{Q_n\})$$

$$= \mathbb{E}\{Q_n\} + \alpha_n (r - \mathbb{E}\{Q_n\}).$$

Consider $|\mathbb{E}\{Q_{n+1}\} - r|$, we have

$$|\mathbb{E}\{Q_{n+1}\} - r| = |\mathbb{E}\{Q_n\} + \alpha_n (r - \mathbb{E}\{Q_n\}) - r| = |(1-\alpha_n)\mathbb{E}\{Q_n\} - (1-\alpha_n) r|$$

$$= |1-\alpha_n| \, |\mathbb{E}\{Q_n\} - r| = |1-\alpha_n|^{n+2} \, |\mathbb{E}\{Q_n\} - r|.$$

When $n \to \infty$, if $\alpha_n = \frac{c}{\sqrt{n}}$, then $0 < \alpha_n < 2$ holds. This yields

$$\lim_{|1-\alpha_n|<1} |1-\alpha_n|^{n+1} \, |\mathbb{E}\{Q_n\} - r| = 0 \Rightarrow \lim_{n\to\infty} |\mathbb{E}\{Q_n\} - r| = 0 \quad \square.$$

Problem 4.

1/ we can formulate by DP with Fixed-Horizon $T = N$ stages correspondy to N days in which we can sell the items.

$\Rightarrow \forall h \in [0, N-1]$, we have:

+/ $S_h \in \mathcal{S}$ be the state r.v. at stage $h \Rightarrow \mathcal{S} = \{0, .., s_0 ; s_0 \in \mathbb{N}\}$
+/ $A_h \in \mathcal{A}$ be the action r.v. at stage h repriant for price $a_h \Rightarrow \mathcal{A} = \mathbb{R}_{\geq 0}$.
+/ $R_{h+1} \in \mathbb{R}_{\geq 0}$ be the revenue reward of A_h at h.

Since we can sell at most 1 item per day with prob. $\lambda(a_h)$, so the state-reward transition kernel $p(R_{h+1}, S_{h+1} | S_h, A_h)$ is

$$\begin{cases} p(R_{h+1} = a_h, S_{h+1} = s_h - 1 \mid S_h = s_h, A_h = a_h) = \lambda(a_h) \\ p(R_{h+1} = 0, S_{h+1} = s_h \mid S_h = s_h, A_h = a_h) = 1 - \lambda(a_h) \end{cases}$$

$\Rightarrow$ The objective function we want to maximize is the cumulative expected return!

$$v_\pi(s_0) = \mathbb{E}_\pi \left\{ \sum_{h=0}^{N} R_{h+1} \mid S = s_0 \right\}.$$

2/ Using the backward recursion, we have the optimal policy $\pi_h^*(a|s) \geq 0$ for all stages are!

$$\begin{aligned} a \in \operatorname*{argmax}_{a' \in \mathcal{A}} & \left\{ \sum_{s', r} p(r, s' | s, a) \left[r + V_{h+1}(s') \right] \right\} \\ &= \operatorname*{argmax}_{a_h \in \mathbb{R}_{\geq 0}} \left\{ \lambda(a_h) \left[a_h + V_{h+1}(s_h - 1) \right] + \left[1 - \lambda(a_h) \right] V_{h+1}(s_h) \right\} \\ &= \operatorname*{argmax}_{a_h \in \mathbb{R}_{\geq 0}} \left\{ \beta e^{-a_h} \left[a_h + V_{h+1}(s_h - 1) \right] + \left(1 - \beta e^{-a_h} \right) V_{h+1}(s_h) \right\} \end{aligned}$$

3, Consider $S_h \geq 0$, let $f(a_h) = \beta e^{-a_h}(a_h + V_{h+1}(S_h - 1)) + (1 - \beta e^{-a_h}) V_{h+1}(S_h)$

since $\Pi_h^*(S_h)$ is the optimal price-setting policy, then

$$\Pi_h^*(S_h) = \arg\max_{a_h \in \mathbb{R}_{h,0}} f(a_h).$$

Take derivative $\frac{\partial f(a_h)}{\partial a_h}$, we have $\frac{\partial f(a_h)}{\partial a_h} = 0$ when

$$e^{-a_h}\{1 + a_h - V_{h+1}(S_h - 1) + V_{h+1}(S_h)\} = 0$$

Since $\ln(e) \neq 0$ & $V_h(S_h)$ are monotically non-decreasing on S_h, i.e., $V_{h+1}(S_h - 1) \leq V_{h+1}(S_h)$, we get

$$a_h = 1 + V_{h+1}(S_h) - V_{h+1}(S_h - 1) \in \mathbb{R}_{\geq 0}.$$

$$\Rightarrow \quad \Pi_h^*(S_h) = 1 + V_{h+1}(S_h) - V_{h+1}(S_h - 1). \quad (8)$$

In addition, we have

$$\begin{aligned} V_h(S_h) &= \max_{a_h \in \mathbb{R}_{\geq 0}} \{\sum_{s',r} p(r, s' | S_h, a_h)(r + V_{h+1}(s'))\} \\ &= \max_{a_h \in \mathbb{R}_{\geq 0}} \{\beta e^{-a_h}\{a_h + V_{h+1}(S_h - 1)) + (1 - \beta e^{-a_h}) V_{h+1}(S_h)\} \\ &= \beta \exp(-1 - V_{h+1}(S_h) + V_{h+1}(S_h - 1)) + V_{h+1}(S_h) \end{aligned}$$

$(8) \Rightarrow = \beta e^{-\Pi^*(S_h)} + V_{h+1}(S_h)$. □ (9)

4a, Since

$$V_h(S_h) = \max_{a_h \in \mathbb{K}_{\geq 0}} \left\{ \beta e^{-a_h} \left[a_h + V_{h+1}(S_h - 1) \right] + (1 - \beta e^{-a_h}) V_{h+1}(S_h) \right\}$$

and

$$V_h(S_h - 1) = \max_{a_h \in \mathbb{K}_{\geq 0}} \left\{ \beta e^{-a_h} \left[a_h + V_{h+1}(S_h - 2) \right] + (1 - \beta e^{-a_h}) V_{h+1}(S_h - 1) \right\}$$

$\Rightarrow$ $\forall\, a_h \in \mathbb{K}_{\geq 0}$, $\beta \in (0, 1)$, we obtain

$$\begin{aligned} V_h(S_h) - V_h(S_h - 1) \geq & \max \{ \beta e^{-a_h} [a_h + V_{h+1}(S_h - 1)] + (1 - \beta e^{-a_h}) V_{h+1}(S_h) \\ & - \beta e^{-a_h} [a_h + V_{h+1}(S_h - 2)] - (1 - \beta e^{-a_h}) V_{h+1}(S_h - 1) \} \\ \geq & \max_{a_h \in \mathbb{K}_{\geq 0}} \{ \beta e^{-a_h} [2V_{h+1}(S_h - 1) - V_{h+1}(S_h) - V_{h+1}(S_h - 2)] + V_{h+1}(S_h) - V_{h+1}(S_h - 1) \} \end{aligned}$$

≥ 0, $\forall h$. $\Rightarrow$ $V_h(S_{h-1}) \leq V_h(S_h)$, $\forall h$,

i.e., $V_h(S_h)$ is indeed non-decreasing on S_h □.

4b, Using the result from (8), we have

$$\Pi_h^2(S_h) = 1 + V_{h+1}(S_h) - V_{h+1}(S_h - 1)$$

and

$$\Pi_h^2(S_h - 1) = 1 + V_{h+1}(S_h - 1) - V_{h+1}(S_h - 2).$$

Since $V_h(S_h)$ is non-decreasing on S_h, i.e.,

$$V_{h+1}(S_h - 2) \leq V_{h+1}(S_h - 1) \leq V_{h+1}(S_h),$$

we obtain

$$\Pi_h^2(S_h - 1) \geq \Pi_h^2(S_h),$$

i.e., $\Pi_h^2(S_h)$ is non-increasing on S_h. □

4 c) Consider the term

$$V_t(S_t) = \max_{a_t \in \mathbb{R}_{\geq 0}} \left\{ \beta e^{-a_t} \left(a_t + V_{t+1}(S_t - 1) \right) + \left(1 - \beta e^{-a_t}\right) V_{t+1}(S_t) \right\}$$

- If $S_t = 0$ $\Rightarrow$ we don't have any item to sell, i.e., $\Pi_t(a_t \mid S_t = 0) = 0$

$\Rightarrow V_t(S_t = 0) = 0$.

- If $S_t \geq N - t$ $\Rightarrow$ we have # items $\geq$ # days we can sell it. Due to each day, we can only sell at most 1 item

$$\Rightarrow V_t(S_t) = \mathbb{E}_{\Pi^*}\left[\sum_{k=t}^{N} R_{k+1} \mid S_t = S_t \right] = (N - t)\, \beta e^{-1}. \quad (10)$$

- If $0 < S_t < N - t$, since (9), we have

$$V_t(S_t) = \beta e^{-\Pi^*(S_t)} + V_{t+1}(S_t),$$

We can set the optimal policy $\Pi_t^*(S_t)$ until $S_t \geq N - t$, combining with the result from (10), we obtain.

$$V_t(S_t) = \sum_{i=t}^{N-t} \beta e^{-\Pi_i^*(S_t)} + S_t \beta e^{-1}.$$

As a result, the closed form solution for V_t is

$$V_t(S_t) = \begin{cases} (N-t)\beta e^{-2} & \text{if } S_t \geq N - t \\ \sum_{i=t}^{N-t} \beta e^{-\Pi_i^*(S_t)} + S_t \beta e^{-1} & \text{if } 0 < S_t < N - t \\ 0 & \text{if } S_t = 0 \end{cases}$$

□.